This is a pre-copyedited, author-produced version of an article accepted for publication in [insert journal title] following peer review. The version of record [insert complete citation information here] is available online at: https://academic.oup.com/bioscience/article/74/9/640/7745614.

Title:

Going Global by Going Local: Impacts and Opportunities of Geographically Focused Data Integration

Authors:

Malgorzata Lagisz (**corresponding author**) – Evolution & Ecology Research Centre and School of Biological, Earth and Environmental Science, University of New South Wales, Sydney, Australia; Theoretical Sciences Visiting Program, Okinawa Institute of Science and Technology Graduate University, Onna, Japan; m.lagisz@unsw.edu.au; ORCID: 0000-0002-3993-6127

Martin Westgate – Atlas of Living Australia, National Collections & Marine Infrastructure, CSIRO, Canberra, Australia; martin.westgate@csiro.au; ORCID: 0000-0003-0854-2034

Dax Kellie – Atlas of Living Australia, National Collections & Marine Infrastructure, CSIRO, Canberra, Australia; dax.kellie@csiro.au; ORCID: 0000-0001-7408-273X

Shinichi Nakagawa – Evolution & Ecology Research Centre and School of Biological, Earth and Environmental Science, University of New South Wales, Sydney, Australia; Theoretical Sciences Visiting Program, Okinawa Institute of Science and Technology Graduate University, Onna, Japan; s.nakagawa@unsw.edu.au; ORCID: 0000-0002-7765-5182

Running title: Geographically focused data integration

Abstract

Biodiversity conservation is a global challenge that requires the integration of global and local data. Expanding global data infrastructures have opened unprecedented opportunities for biodiversity data storage, curation, and dissemination. Within one such infrastructure – the Global Biodiversity Information Facility (GBIF) – these benefits are achieved by aggregating data from over 100 regional infrastructure 'nodes'. GBIF nodes play a critical role by building and maintaining collaborations that support data sharing and reuse, while also fostering local research, traditional knowledge, innovation, and engagement. Therefore, regional-scale biodiversity infrastructures benefit scientific communities in ways that exceed their core function of contributing to global data aggregation; yet these additional scientific impacts are rarely quantified. To fill this gap, we characterise the scientific impact of the Atlas of Living Australia (ALA), one of the oldest and largest GBIF nodes, as a case study of a regional biodiversity information facility. Our discussion reveals the multi-faceted impact of the regional biodiversity data infrastructure. We showcase the global importance of such infrastructures, datasets, and collaborations.

Keywords

bibliometrics, data infrastructures, global impact, information dissemination, meta-research

Introduction

The ongoing sixth mass extinction is a global challenge (Ceballos and Ehrlich 2018, Cowie et al. 2022). Tackling this global biodiversity crisis requires cross-national efforts informed by multi-dimensional data (Bisby 2000, Franklin et al. 2017, Heberling et al. 2021). Such data is usually gathered locally, consisting of information like species occurrences, species characteristics, environment and land use, pieced together into "big data" global datasets and hosted on specialised data infrastructures: e.g., Encyclopedia of Life (eol.org), Ocean Biodiversity Information System (obis.org), International Barcode of Life (ibol.org), Global Biodiversity Information Facility (gbif.org) (Kays and Wikelski 2023). Global datasets and infrastructure have enabled remarkable insights into the current, past, and predicted state of biodiversity changes (e.g., Warren et al. 2018, Staude et al. 2020, Callaghan et al. 2021).

Global-level data and analyses are crucial to conservation, yet they are insufficient to protect biodiversity on their own (Armitage et al. 2020). Achieving global biodiversity protection requires knowledge coproduction and engagement that is socially and environmentally just – a task that cannot be achieved using a global "one-size-fits-all" approach. Instead, this knowledge must be built locally by developing relationships of trust and meaningful participation (Raymond et al. 2022). The importance of integrating knowledge from both established monitoring programs *and* local and indigenous community knowledge is recognised as a global conservation priority, for example Target 19 of the Kunming-Montreal Global Biodiversity Framework (Biosafety Unit, 2023). Local and regional biodiversity infrastructures are essential in facilitating this process (Canhos et al. 1993, 2015, Tauch and Al-Dilaimi 2019, Schulman et al. 2021, Urban et al. 2022, Peterson et al. 2023). Any characterisation of local and regional infrastructures, however, is rare.

It is necessary to conduct in-depth impact assessment of regional biodiversity platforms separately from the assessments of the global biodiversity infrastructures because the latter may underplay or miss local impacts or outputs produced by non-academic stakeholders. For example, a recent study's assessment of 4,000 academic studies linked to Global Biodiversity Information Facility (GBIF), one of the largest open-source data infrastructures of global biodiversity data, did not include non-academic publications or non-publication outputs (Heberling et al. 2021).

Although GBIF aggregates over two billion species occurrence records, it is critical to note that much of this data comes from local and regional "node" organisations (Figure 1), which in turn collect and aggregate contributions from hundreds of local data providers. A number of these node organisations have become regional biodiversity and collaboration hubs, bringing unique opportunities and benefits (Edwards 2004). These benefits can include direct engagement with not only regional data providers but also researchers, practitioners, policymakers, and the public. This opportunity is achieved by building customised online platforms, digitising local specimen collections, close and personalised collaboration with local agencies and institutions, providing training and support, and engaging local citizen scientists and Indigenous people via collaborative projects (Brenton et al. 2018, Roger et al. 2023). Regional hubs are thus critical to aggregating standardised data that allow GBIF to facilitate the global biodiversity syntheses (Heberling et al. 2021), but they receive limited attention on the global stage.

A growing worldwide community of regional biodiversity infrastructures, known as "Living Atlases", has been enabled by the Atlas of Living Australia (ALA) (Atlas of Living Australia 2022). While the majority of the 106 GBIF nodes pass data directly to GBIF, and do not

maintain local data portals, the Living Atlases (https://living-atlases.gbif.org) use the ALA open-source codebase to build their own data and systems before federating to the global database. This enables them to collate, curate and link data that are specific to their own regions, including environmental variables (spatial and historical), species profiles, images and sounds, trait measurements, conservation status, and taxonomic information (Belbin et al. 2021). The aim of this organisational model is to improve the value of the living atlases as resources for addressing regional-level research, conservation and policy-related questions. Given these aims, as well as the ALA's enabling role in developing the 'living atlas' model, it is valuable to ask to what extent the ALA has delivered direct scientific and community impact at the regional scale, in addition to its role as a node for GBIF.

Assessing the impact of biodiversity data platforms rarely extends beyond their contents or functionality of the underlying digital infrastructure (e.g. Vattakaven et al. 2016, Dias et al. 2017, Schulman et al. 2021). Thus, little is known about how unique data from such regional biodiversity information hubs translate into specific outputs, such as peer-reviewed journal articles (and their traditional measures of impact), non-peer-reviewed outputs (e.g., reports, books, software, blogs – "grey outputs", produced by diverse stakeholders of such platforms), and how these outputs influence our knowledge and lives (e.g., via patents, policies, information platforms, social media). We need to know how digital biodiversity infrastructure can foster trust, collaborations, and broader scientific and social impact. To achieve this, we should go beyond counting data records or academic articles. We must look at a broader range of output types and influences, the key players, and collaborations. Such an approach can support the idea that decentralized approaches are critical to maximising

the societal impact of the biodiversity data (Sterner et al. 2020) and, ultimately, offer recognition to the role of regional initiatives in resolving global issues.

Our objective is to provide an in-depth and comprehensive overview the impacts of the diverse ALA-related outputs on the scientific and non-scientific communities. We address this main objective by characterising ALA-related outputs to answer four specific questions: 1) What can ALA-related academic and grey outputs tell us about the impacts of regional biodiversity infrastructures and data? 2) What can we learn from the ALA about local and global collaborations and author diversity? 3) Are the ALA-related outputs accessible to a broad readership? 4) How impactful is the academic research facilitated by the ALA?

Gathering key information about the ALA-related outputs

The ALA was established in 2010 by the Australian Government's National Collaborative Infrastructure Strategy (NCRIS) and is hosted by The Commonwealth Scientific and Industrial Research Organisation (CSIRO). It aggregates species occurrence records from government monitoring programs, research projects, museums, herbaria, and popular citizen science apps like iNaturalist (Di Cecco et al. 2021, Campbell et al. 2023). Via GBIF, it feeds species occurrence records from Australia into the global system. The list of ALA-related research outputs used in this project was manually curated over several years by ALA employees. They set up publication alerts on the main academic publishing platforms using ALA-related keywords, DOI, and records linked to resources on ALA platform. The list of ALA-related research outputs is stored as a private shared library on Zotero reference management platform and updated weekly. The Zotero database contains basic bibliographic information

on the collected outputs, such as: output type, title, authors, publication source (e.g., journal name), publisher, ISBN, ISSN, DOI, url, abstract, date, and automatic keywords. In the database, manual tags are added by ALA curators to indicate how a given item engaged with or related to the ALA: whether specific ALA-provided resources were used (e.g., research data or services), cited, or informally mentioned. When ALA resources and services were used directly, the type of the resource is also coded as: species occurrences records, species inventory lists, species profiles, climatic data, spatial portal, and modelling services. Supplementary Material files contain additional details on data collation and coding. For analyses, we exported the list of 3,866 bibliographic records from Zotero on 30 March 2023 as a .csv file. We used these bibliographic records as a starting point for supplementing the records with additional data from other sources to allow for deeper analyses on the impacts of the ALA platform. Specifically, we extracted related data from publicly available datasets and information services (Unpaywall, Journal Citation Records, Scimago, OpenAlex, Altmetrics), as detailed in the Supplementary Material 1. In brief, out of 3,866 items in the original Zotero dataset we removed 92 published in early 2023, leaving 3,774 for our analyses. We extracted and processed authorship information of the ALA impact observations. The 3,774 ALA-related outputs had a total of 15,395 individual authors recorded, with 9,813 unique combinations of family names and initials. We found country of affiliation information for 8,672 authors of 1,765 journal articles from OpenAlex. We gathered data on the open access status of 1,811 journal articles from Unpaywall. We then retrieved academic citations counts (Crossref) for 2,039 journal articles, and Altmetric attention scores (Altmetrics) for 1,519 journal articles. We annotated 1,543 journal articles with information on journal Impact Factors from Journal Citation Records. Finally, for 1,886 articles, we found SCImago journal ranking position, country, region, and subject category

information. Additionally, we used GBIF API to conduct a comparison of citations that Australian datasets get via ALA vs GBIF.

The patterns in ALA impact observations

ALA impact observations list captured predominantly peer-reviewed publications. Of 3,774 items published in 2007–2022, 2,406 (70%) were classified as journal articles, of which 2,070 had associated DOI. Among the other 15 item types (thereafter "grey outputs", totalling 1,358 records), only 98 conference papers had DOI. Lack of DOI made it hard to reliably integrate information from other data platforms, which reduced the scope of our analyses of grey outputs. In terms of linguistic diversity, we were limited by lack of reliable meta-data. Where publication language was specified, English was the dominant language (>99%), with only 13 documents in other languages (Spanish, Portuguese, French, German, Russian, Indonesian, Italian, Swedish, and Chinese).

Comparing the scope of journal articles and grey outputs

Our custom-curated dataset revealed the number of ALA-related outputs steadily increased until around 2017, when direct citations of ALA appear to shift toward citations via GBIF (Figure 2A; notably, in 2023, only 32.7% of citations of Australian data were via ALA, with the rest coming from GBIF). This growing trend in citations over the years is mainly driven by publication rates of journal articles (thereafter "articles", 2,406 records), which dominate our data set (Figure 2B). Other types of ALA-related outputs (government, organisation, media and related items, i.e. "grey outputs") generally also increased over time, but their

citations are harder to track and thus are not directly comparable. Next, we compared articles and grey outputs regarding the ways of use / crediting of the ALA resources, and differences in thematic focus between journal articles and grey output types, resulting in four key observations.

First, journal articles are twice as likely to directly use specific ALA-provided resources (Figure 3A), such as research data or services. They also tend to explicitly cite ALA-provided resources more often. On the other hand, grey outputs seem to informally mention or discuss ALA in their text more often than are journal articles.

Second, when ALA resources are used directly, all six main categories of data and services available via ALA are represented in both output types (Figure 3B), with species occurrences records being the main type of utilised resource. However, grey outputs tend to use more often the "core data", such as species occurrence records and specific species lists, rather than more advanced services such as digital maps and models.

Third, around half of the titles and abstracts contained words that indicate that they study Australian species or ecosystems (Figure 3C). Such words included names of Australian states, regions, cities, and commonly studied endemic and charismatic species (e.g., koala, platypus, kangaroo). Outputs published as journal articles tended to mention these words more frequently than grey outputs.

Fourth, a small proportion of titles and abstracts contained words indicating an involvement of the broader community in the research, for example, via citizen science, volunteering

and/or data crowdsourcing (Figure 3D). This proportion was comparable in journal articles and in the grey outputs.

Authorship diversity and collaboration

Around 8% of 9,813 unique individual authors authored both journal articles and grey outputs. Notably, the pool of unique individual authors of grey outputs (2,742) is much smaller than the pool of authors of journal articles (7,827). Also, 11% of the grey outputs either had no authorship captured or had a record of a group authorship only (e.g., name of an institution, organisation, or community). This may have influenced the observation that, on average, grey outputs typically have smaller author team sizes (Supplementary File).

Over the years, the size of the individual authorship teams increased for journal articles. In contrast, for 1,212 records of grey outputs with listed individual authors, no clear trend for team size could be observed. Yet, "Big Team" science outputs started appearing in the last decade – we found 11 journal articles with more than 50 authors, and one such grey output (with 211 individually-named authors; (Royal Botanic Gardens, Kew 2020)).

Affiliated country of journal article authors

Analyses of author country of affiliations tells us whether infrastructure is used globally and whether research teams are international. The 2,681 authors of 1,765 journal articles represented 54 different countries. Australia (67.8%) dominated the authorship of journal articles, followed by the USA (8.0%), Germany (3.8%), and the United Kingdom (2.3%) (Figure 4). The majority of these journal articles (70.0%) had all authors affiliated with the same country (usually Australia – 58%). For articles with authors affiliating with different

countries, the most common collaborations of the Australian authors involved USA, Germany, New Zealand, and UK (Supplementary File).

Publication patterns of journal articles

Analyses on publication journals can approximate geographical and disciplinary reach of the produced knowledge. ALA-related articles were published in 753 different journals, including 413 journals with calculated Impact Factors. Twenty of the ALA-related journal articles were published in 10 different journals with Impact Factor larger than 15, mostly in Nature Communications, Nature Ecology and Evolution, and Science (Figure 5A). Seven of those articles explicitly used the ALA-collated resources, such as species occurrence records, another seven only mentioned the ALA, three had ALA-affiliated authors and the remaining three cited the ALA. However, the majority of the articles in our data set tended to be published in journals with much lower Impact Factors, with the median value at 3.3. The ten most popular journals included cross-disciplinary journals like PLOS ONE and Zootaxa, as well as Australia-focused journals like Austral Ecology and Austral Entomology (Figure 5B). Australia-focused journals tend to be of lower impact, as indicated by the under-representation of the Pacific region in the top quartile (Q1) of the SCImago Journal & Country Rank data (Figure 5C). Still, the majority (62%; 1,162 out of 1,886) of the ALA-related articles were published in the top quartile (Q1) journals, usually associated with Western Europe and North American readership.

Accessibility of journal articles

For effective biodiversity conservation, it is critical that produced knowledge is freely accessible to broad readership. Overall, more than half (63% of the articles with known open access status) of the journal articles related to ALA have an Open Access status allowing free access to the full article contents (Figure 6A). Gold open-access publishing model was the most popular, followed by Green open access with preprints deposited to public repositories such as bioRxiv and EcoEvoRxiv. Publishing openly accessible articles was embraced early on by the authors and remained high over the last decade (Figure 6B).

Impact of journal articles

Citation counts are a traditional quantitative measure of research impact. Most ALA-related journal articles were cited by other academic publications at least once, with a strongly right-skewed distribution of citations (Supplementary File). While the majority of the articles received less than 10 citations (median = 6), there were also 45 articles with more than 100 citations, and 4 articles with more than 500 citations. The latter group included three methodological articles focused on predicting species distributions using Maxent models (Elith et al. 2011, Guisan et al. 2013, Philips et al. 2017). In these articles, the ALA was mentioned as an example of implementing Maxent models to facilitate the use of Maxent for researchers and the broader public. The fourth article (Wieczorek et al. 2012) is dedicated to Darwin Core, a widely used data standard for biodiversity information, noting the ALA as an example of successful data integration and dissemination using this standard.

ALA-related journal articles were also often mentioned outside the traditional academic publications. Altmetric score (officially called Altmetric Attention Score; altmetric.com) is a weighted count of the amount of attention received by an article from multiple categories of information sources. Out of 1,519 articles with Altmetric data, 79 had Altmetric scores higher than 100, and 4 higher than 1,000, indicating extreme interest. These articles tended to be published in top-ranked journals, such as Proceedings of the National Academy of Sciences of the USA (two articles), and Proceedings of the Royal Society B: Biological Sciences (one article). They also tackled broad questions, such as past and future declines of biodiversity (Cazzolla Gatti et al. 2022, Roycroft et al. 2021), or changes in populations of specific species of high public interest (Bino et al. 2019, Crates et al. 2021). Altmetric scores adjusted for journal context show that most ALA-related articles receive more attention than other articles from the same journals (Supplementary File), revealing significant impact beyond the top-ranked journals. Data on the counts broken into individual public attention components showed that ALA-related articles are cited in policies (76 articles), patents (4), and Wikipedia (280). Attention by news sources was substantial, comprising 3,334 mentions for 291 articles. Seven articles received over 100 news mentions. Most of those articles were among the set of articles with the highest Altmetric attention scores, indicating a cascading effect of the media stories on the other types of public attention, such as social media. The social media attention included mentions on Twitter (28,730 mentions for 1,387 articles), Facebook (848 mentions for 452 articles), and blogs (660 mentions for 347 articles).

The implications of the ALA impact assessment for biodiversity research

Beyond collating and hosting regionally relevant data, we assessed The Atlas of Living Australia publication output, research collaborations and public engagement locally and globally. This can be seen in the answers to our four specific questions we discuss in turn below.

1. What can ALA-related academic and grey outputs tell us about the impacts of regional biodiversity infrastructures and data?

We found that the list of ALA-related outputs is dominated by journal articles, accounting for 64% of the curated list of items. Journal articles appear to drive the increase in the number of captured outputs across the years, and often directly use or cite data and services provided by the ALA. In contrast, grey outputs are more likely to mention or discuss the ALA. This result is not surprising, given that journal articles are usually research-oriented and follow formalised citing conventions, while many of the grey outputs offer secondary recounts of research findings or initiatives in formats that are less formalised but accessible to the broader public, such as magazine articles, blogs, and videos (Pappas and Williams 2011).

We observed that both articles and grey outputs use the whole range of resources and services provided by the ALA. The georeferenced species occurrence records and species distribution maps are the core of the ALA and enabled 71% of the use cases. Climate data, species lists, and spatial and modelling portals comprised 26% of the use cases. The resource with the least citations was the species profile database (final 3%), likely because

these are brief summary pages similar to Wikipedia pages. In line with our expectation, the majority of the outcomes focus on the Australian species or ecosystems, given that at least half mention some of the Australia-related words we searched for in the titles and abstracts. A small but significant portion (around one-tenth) of both journal articles and grey outputs can be linked to community contributions, clearly reflecting ALA's role in connecting citizen scientists, volunteers, local organisations, and researchers (Belbin and Williams 2016, Brenton et al. 2018).

2. What can we learn from the ALA about local and global collaborations and author diversity?

We found that almost one in ten of the captured individual authors contributed to both journal articles and other output types. Such versatile authors are important for knowledge transfer to other stakeholders and for forming broader collaborations. Overall, the authorship teams were very diverse in size and level of collaboration. We observed a subset of outputs connected very weakly via authorship, but also a large core of the outputs that share at least some of their authors, forming collaborative networks (Supplementary File). For the journal articles, the historical increase in authorship team sizes indicates the increasing importance of large-scale collaborative efforts (Cheruvelil and Soranno 2018). Such efforts can be linked with the use of Australian biodiversity data in regional studies, but also in global studies. It may also reflect the global trend of increased co-authorship (Thelwall and Maflahi 2022). The broad network of international collaborations includes authors affiliated with over 50 countries. However, these connections (and publication volumes) were dominated by developed English-speaking countries, such as USA and UK. This observation is in line with the pattern of the strong geographic imbalance in publishing

patterns across disciplines (King 2004), observed also within disciplines, e.g., in ecological (Tydecks et al. 2018) and climate-science research (Tandon 2021), (but see Miao et al. 2022).

3. Are the ALA-related outputs accessible to a broad readership?

More than half (63%) of ALA-related journal articles is currently openly available as full text under gold, brown, green or hybrid open-access schemes. This is a higher proportion than that reported for GBIF-linked academic articles (30% overall, 38% between 2016 and 2019; Heberling et al. 2021). Providing open access to the full contents of the articles (and other outputs) is essential for enabling broad dissemination, assessment, and applications of the research findings of individual studies (Bourne et al. 2008). Differences between where ALA-linked and GBIF-linked articles are most frequently published might reflect differences in open-access priorities in journals that publish local or global research studies.

4. How impactful is the academic research facilitated by the ALA?

The picture of the impact of ALA-related journal articles is complex but far from unexpected. Traditional metrics of academic impact include citation counts for individual articles, publication journal impact factor and journal rank quartile. While the majority of the analysed journal articles gathered modest citation numbers, which may be partially explained by many having recent publication dates, a small subset of articles received outstanding numbers of citations, characteristic of seminal publications (Herrmannova et al. 2018). These articles introduced or elaborated on methodological innovations of global interest, and they were published in journals with high impact factor values and broad

readership. Most articles with fewer citations also made it into top-ranked journals, potentially reflecting the perceived importance or quality of the work. Many of the most popular publication journals appear to focus on research directly related to Australia, such as characterising local species or local species distributions, which may result in smaller readership base and lower citation rates. Nevertheless, such publications may be critical to local conservation, planning and policy making. For example, the ALA has been used in high-profile articles on the impact of invasive diseases on ecological niches (Scheele et al. 2023), long-term data collection of koala individuals through citizen science (Danaher et al. 2023), and is a resource of indigenous ecological knowledge (Belbin et al. 2021). Similarly, the regional classification of the journals matched the results of the keyword analyses on the article titles and abstracts, indicating strong regional focus of the articles, in line with the core mission of the ALA (Blackburn et al. 2014, Belbin et al. 2021). Notably, the ALA-related outputs are increasingly mentioned in the policies, patents, Wikipedia pages, news outlets and social media; it attests to the local and global influence and interest these outputs generate. The far-reaching influence reflects the sustained funding of the ALA from the Australian Government through the National Collaborative Research Infrastructure Strategy.

Potential limitations of our assessment

Our work should be interpreted with six limitations in mind. First, the collected body of the ALA-related outputs contains few items in languages other than English. These outputs are likely underrepresented and underestimated due to the searches being conducted using English-language platforms and keywords (Chowdhury et al. 2022), as well as frequently missing information on the language of the item itself. Second, other information was often missing from the bibliographic records, such as DOI, author affiliations, and abstracts, and

author affiliations, especially for the grey outputs (e.g., 14% of journal articles did not have an abstract, vs. 44% for grey outputs). This issue has limited the scope of possible data integration from other sources and, ultimately, the range of analyses and the reliability of the insights we could provide in this work. Third, we have likely underestimated Australian context and community involvement by using a set of predetermined keywords to text mine titles and abstracts of the ALA-related outputs, which was additionally hindered by missing abstracts. Fourth, our analyses of author collaboration networks could be inaccurate when author names are used as a proxy for author identity due to potential inconsistencies in recorded author names (same person with different name variants), as well as the name overlaps (different people with the same name). Also, 11% of grey outputs did not attribute authorship to individuals, implying organisations or communities as the authors, which likely affected the accuracy of our co-authorship analyses. Fifth, Altmetric scores measure attention, not true impact, and come with many inherent limitations that warrant caution in the interpretation of Altmetric-based indices (Konkiel 2016). Sixth, the grey outputs could not be assessed in the same way as journal articles in terms of their impact, as Altmetric scores, citation counts, and impact factors, are not trackable for them.

Future directions and perspectives for impact assessment of biodiversity infrastructures

Analyses of various outcomes and their impacts are always challenging due to the effort needed for extensive data gathering, processing and analyses (Brenninkmeijer 2022). Such projects are likely to deal with inconsistent or incomplete datasets and disparate sources of information that are not harmonized or are accessible only under a subscription model. The challenges related to data quality cannot be mitigated simply by the manual curation and annotation of the records, as this task is not feasible for high data volumes. Better data

quality requires systemic changes. For example, the DOI is a universal and the most reliable digital output identifier when incorporating information from multiple sources, and authors of grey outputs should ensure that they have one. Similarly, author ORCID numbers are a simple and reliable way of recording author identities, and their universal use would greatly aid authorship analyses. ORCID is free and easy to register for, thus, is potentially suitable even for non-academic authors. Author affiliations are notoriously messy in the bibliographic records, which could be resolved by the universal adaptation and reporting of machine-readable author affiliation metadata standards, including GRID institutional identifiers (Aspeslagh et al. 2021). It is important to consider the impact or received attention of individual outputs beyond traditional metrics of citations or impact factors, especially for the local-scale impacts. For example, by collecting engagement data and feedback from the stakeholders, such as data collectors and users, with special attention to the indigenous and underrepresented communities (Valdez et al. 2023). Also, we should take into account disciplinary or linguistic differences when interpreting any indices. Finally, freely accessible bibliometric information can be used for impact evaluations via platforms such as OpenAlex, Unpaywall and Crossref. In our work, we have combined data from these platforms with other openly available datasets to build a reproducible bibliometric research workflow circumventing commercial interests. Similar workflows can facilitate multifaceted impact assessments of research-related outputs across all subject areas, beyond ecology.

In terms of the contributions to our understanding of practices related to global ecology, we stress that global research conservation cannot be achieved without examining how we collect, use, and disseminate data locally. Thus, our work highlights new approaches to examining outputs and impacts related to regional bioinformatics infrastructures. The ALA,

itself, appears to live up to its ambitions by contributing growing and high-quality body of journal articles and other types of outputs, bridging national and international collaborations, and fostering social impacts beyond academia.

Conclusions

Overall, this work exemplifies how both academic and the social impacts can be inferred from the body of outputs linked to a regional bioinformatics infrastructure. Our findings show how such regional infrastructure is uniquely positioned to enable discoveries, collaborations, and broader engagement essential to human societies and the conservation of biodiversity. Thus, supporting over 9,000 authors and 2,406 scientific publications (up to 2023), the ALA delivers trusted biodiversity data services for Australia, supporting world-class research and decision-making, with the impacts extending beyond the Australian territory.

Data Availability

The study analysis was based on a collated outputs dataset provided by the Atlas of Living Australia. Additional information was fetched from online databases or downloaded from websites, as outlined in the Methods section. The data sets and code supporting the results of this article are available on the GitHub (https://github.com/mlagisz/ALA_research_weaving) and archived in Zenodo repository

(https://zenodo.org/record/8024192; DOI: 10.5281/zenodo.8024191 [link to be updated for the accepted version]).

Abbreviations

ALA: Atlas of Living Australia; API: Advanced Programming Interface; DOI: Digital Object Identifier; GBIF: Global Biodiversity Information Facility; IF: Impact Factor; SE: Standard Error; UK: United Kingdom; USA: United States of America.

Acknowledgments

The authors would like to acknowledge the ALA employees for collating the database of ALA-related outputs and annotating the records with manual tags: Corinna Paeper, Ely Wallis, and Olivia Torresan. Callum Waite wrote code for comparing the ALA & GBIF citation databases. This work has been funded by the Commonwealth Scientific and Industrial Research Organisation (CSIRO) / Commonwealth Government Contract (RG213588). MW and DK are employees of the Atlas of Living Australia.

References

Armitage D, Mbatha P, Muhl E-K, Rice W, Sowman M. 2020. Governance principles for community-centered conservation in the post-2020 global biodiversity framework. Conservation Science and Practice 2: e160.

Aspeslagh P, Engels T, Guns R. 2021. Proceedings of the ICTeSSH 2021 conference. Paper presented at Proceedings of the ICTeSSH 2021 conference. 28 June 2021.

Atlas of Living Australia. 2022. Atlas of Living Australia Year in Review 2021–22.

Belbin L, Wallis E, Hobern D, Zerger A. 2021. The Atlas of Living Australia: History, current state and future directions. Biodiversity Data Journal 9: e65023.

Belbin L, Williams KJ. 2016. Towards a national bio-environmental data facility: experiences from the Atlas of Living Australia. International Journal of Geographical Information Science 30: 108–125.

Bino G, Kingsford RT, Archer M, Connolly JH, Day J, Dias K, Goldney D, Gongora J, Grant T, Griffiths J, Hawke T, Klamt M, Lunney D, Mijangos L, Munks S, Sherwin W, Serena M, Temple-Smith P, Thomas J, Williams G, Whittington C. 2019. The platypus: evolutionary history, biology, and an uncertain future. Journal of Mammalogy 100: 308–327.

Biosafety Unit. 2030 Targets (with Guidance Notes). (10 March 2024; https://www.cbd.int/gbf/targets).

Bisby FA. 2000. The Quiet Revolution: Biodiversity Informatics and the Internet. Science 289: 2309–2312.

Blackburn R, La SJ, Doherty P. 2014. Revolutionising the Mapping of nature: The atlas of living Australia. AQ - Australian Quarterly 85: 19–23.

Bourne PE, Fink JL, Gerstein M. 2008. Open access: taking full advantage of the content. PLoS computational biology 4: e1000037.

Brenninkmeijer J. 2022. Achieving societal and academic impacts of research: A comparison of networks, values, and strategies. Science and Public Policy 49: 728–738.

Brenton P, von Gavel S, Lecoq M-E, Vogel E. 2018. Infrastructure enabling citizen science – an information supply chain perspective. In: London, editor/s. Citizen Science – Innovation in Open Science, Society and Policy (eds A. Bonn, M. Haklay, S. Hecker, A. Bowser, Z. Makuch & J. Vogel). UCL Press; 2018. 17.

Callaghan CT, Nakagawa S, Cornwell WK. 2021. Global abundance estimates for 9,700 bird species. Proceedings of the National Academy of Sciences of the United States of America 118: e2023170118.

Campbell CJ, Barve V, Belitz MW, Doby JR, White E, Seltzer C, Di Cecco G, Hurlbert AH, Guralnick R. 2023. Identifying the identifiers: How iNaturalist facilitates collaborative, research-relevant data generation and why it matters for biodiversity science. BioScience 73: 533–541.

Canhos DAL, Sousa-Baena MS, de Souza S, Maia LC, Stehmann JR, Canhos VP, De Giovanni R, Bonacelli MBM, Los W, Peterson AT. 2015. The Importance of Biodiversity E-infrastructures for Megadiverse Countries. PLoS biology 13: e1002204.

Canhos VP, Canhos DL, de Souza S. 1993. Establishment of a computerized biodiversity/biotechnology network: the Brazilian effort. Journal of Biotechnology 31: 67–73.

Cazzolla Gatti R, Reich PB, Gamarra JGP, Crowther T, Hui C, Morera A, Bastin J-F, de-Miguel S, Nabuurs G-J, Svenning J-C, Serra-Diaz JM, Merow C, Enquist B, Kamenetsky M, Lee J, Zhu J, Fang J, Jacobs DF, Pijanowski B, Banerjee A, Giaquinto RA, Alberti G, Almeyda Zambrano AM, Alvarez-Davila E, Araujo-Murakami A, Avitabile V, Aymard GA, Balazy R, Baraloto C, Barroso JG, Bastian ML, Birnbaum P, Bitariho R, Bogaert J, Bongers F, Bouriaud O, Brancalion PHS, Brearley FQ, Broadbent EN, Bussotti F, Castro da Silva W, César RG, Češljar G, Chama Moscoso V, Chen HYH, Cienciala E, Clark CJ, Coomes DA, Dayanandan S, Decuyper M, Dee LE, Del Aguila Pasquel J, Derroire G, Djuikouo MNK, Van Do T, Dolezal J, Đorđević IĐ, Engel J, Fayle TM, Feldpausch TR, Fridman JK, Harris DJ, Hemp A, Hengeveld G, Herault B, Herold M, Ibanez T, Jagodzinski AM, Jaroszewicz B, Jeffery KJ, Johannsen VK, Jucker T, Kangur A, Karminov VN, Kartawinata K, Kennard DK, Kepfer-Rojas S, Keppel G, Khan ML, Khare PK, Kileen TJ, Kim HS, Korjus H, Kumar A, Kumar A, Laarmann D, Labrière N, Lang M, Lewis SL, Lukina N, Maitner BS, Malhi Y, Marshall AR, Martynenko OV, Monteagudo Mendoza AL, Ontikov PV, Ortiz-Malavasi E, Pallqui Camacho NC, Paquette A, Park M, Parthasarathy N, Peri PL, Petronelli P, Pfautsch S, Phillips OL, Picard N, Piotto D, Poorter L, Poulsen JR, Pretzsch H, Ramírez-Angulo H, Restrepo Correa Z, Rodeghiero M, Rojas Gonzáles RDP, Rolim SG, Rovero F, Rutishauser E, Saikia P, Salas-Eljatib C, Schepaschenko D, Scherer-Lorenzen M, Šebeň V, Silveira M, Slik F, Sonké B, Souza AF, Stereńczak KJ, Svoboda M, Taedoumg H, Tchebakova N, Terborgh J, Tikhonova E,

Torres-Lezama A, van der Plas F, Vásquez R, Viana H, Vibrans AC, Vilanova E, Vos VA, Wang H-F, Westerlund B, White LJT, Wiser SK, Zawiła-Niedźwiecki T, Zemagho L, Zhu Z-X, Zo-Bi IC, Liang J. 2022. The number of tree species on Earth. Proceedings of the National Academy of Sciences 119: e2115329119.

Ceballos G, Ehrlich PR. 2018. The misunderstood sixth mass extinction. Science (New York, N.Y.) 360: 1080–1081.

Cheruvelil KS, Soranno PA. 2018. Data-Intensive Ecological Research Is Catalyzed by Open Science and Team Science. BioScience 68: 813–822.

Chowdhury S, Gonzalez K, Aytekin MÇK, Baek S-Y, Bełcik M, Bertolino S, Duijns S, Han Y, Jantke K, Katayose R, Lin M-M, Nourani E, Ramos DL, Rouyer M-M, Sidemo-Holm W, Vozykova S, Zamora-Gutierrez V, Amano T. 2022. Growth of non-English-language literature on biodiversity conservation. Conservation Biology: The Journal of the Society for Conservation Biology 36: e13883.

Cowie RH, Bouchet P, Fontaine B. 2022. The Sixth Mass Extinction: fact, fiction or speculation? Biological Reviews of the Cambridge Philosophical Society 97: 640–663.

Crates R, Langmore N, Ranjard L, Stojanovic D, Rayner L, Ingwersen D, Heinsohn R. 2021. Loss of vocal culture and fitness costs in a critically endangered songbird. Proceedings of the Royal Society B: Biological Sciences 288: 20210225.

Danaher M, Schlagloth R, Hewson M, Geddes C. 2023. One Person and a Camera: a relatively non-intrusive approach to Koala citizen science. Australian Zoologist 43: 52–66.

Di Cecco GJ, Barve V, Belitz MW, Stucky BJ, Guralnick RP, Hurlbert AH. 2021. Observing the Observers: How Participants Contribute Data to iNaturalist and Implications for Biodiversity Science. BioScience 71: 1179–1188.

Dias D, Baringo Fonseca C, Correa L, Soto N, Portela A, Juarez K, Tumolo Neto RJ, Ferro M, Gonçalves J, Junior J. 2017. Repatriation Data: More than two million species occurrence records added to the Brazilian Biodiversity Information Facility Repository (SiBBr). Biodiversity Data Journal 5: e12012.

Edwards JL. 2004. Research and Societal Benefits of the Global Biodiversity Information Facility. BioScience 54: 485–486.

Elith J, Philips SJ, Hastie T, Dudík M, Chee YE, Yates CJ. 2011. A statistical explanation of MaxEnt for ecologists. Diversity and Distributions 17: 43–57.

Franklin J, Serra-Diaz JM, Syphard AD, Regan HM. 2017. Big data for forecasting the impacts of global change on plant communities. Global Ecology and Biogeography 26: 6–17.

Guisan A, Tingley R, Baumgartner JB, Naujokaitis-Lewis I, Sutcliffe PR, Tulloch AIT, Regan TJ, Brotons L, McDonald-Madden E, Mantyka-Pringle C, Martin TG, Rhodes JR, Maggini R, Setterfield SA, Elith J, Schwartz MW, Wintle BA, Broennimann O, Austin M, Ferrier S, Kearney MR, Possingham HP, Buckley YM. 2013. Predicting species distributions for conservation decisions. Ecology Letters 16: 1424–1435.

Heberling JM, Miller JT, Noesgaard D, Weingart SB, Schigel D. 2021. Data integration enables global biodiversity synthesis. Proceedings of the National Academy of Sciences 118: e2018093118.

Herrmannova D, Patton R, Knoth P, Stahl C. 2018. Do Citations and Readership Identify Seminal Publications? Scientometrics 115(1): 239–262.

Kays R, Wikelski M. 2023. The Internet of Animals: what it is, what it could be. Trends in Ecology & Evolution 38: 859–869.

King DA. 2004. The scientific impact of nations. Nature 430: 311–316.

Konkiel S. 2016. Altmetrics: diversifying the understanding of influential scholarship. Palgrave Communications 2: 1–7.

Miao L, Murray D, Jung W-S, Larivière V, Sugimoto CR, Ahn Y-Y. 2022. The latent structure of global scientific development. Nature Human Behaviour 6: 1206–1217.

Pappas C, Williams I. 2011. Grey Literature: Its Emerging Importance. Journal of Hospital Librarianship 11: 228–234.

Peterson JD, Kasperowski D, van der Wal R. 2023. Bringing Together Species Observations: A Case Story of Sweden's Biodiversity Informatics Infrastructures. Minerva 61: 265–289.

Philips SJ, Anderson RP, Dudík M, Schapire RE, Blair ME. 2017. Opening the black box: an open-source release of Maxent. Ecography 40: 887–893.

Raymond CM, Cebrián-Piqueras MA, Andersson E, Andrade R, Schnell AA, Battioni Romanelli B, Filyushkina A, Goodson DJ, Horcea-Milcu A, Johnson DN, Keller R, Kuiper JJ, Lo V, López-Rodríguez MD, March H, Metzger M, Oteros-Rozas E, Salcido E, Sellberg M, Stewart W, Ruiz-Mallén I, Plieninger T, van Riper CJ, Verburg PH, Wiedermann MM.

2022. Inclusive conservation and the Post-2020 Global Biodiversity Framework: Tensions and prospects. One Earth 5: 252–264.

Roger E, Kellie D, Slatyer C, Brenton P, Torresan O, Wallis E, Zerger A. 2023. Open Access Research Infrastructures are Critical for Improving the Accessibility and Utility of Citizen Science: A Case Study of Australia's National Biodiversity Infrastructure, the Atlas of Living Australia (ALA). Citizen Science Theory and Practice 8(1): 56.

Royal Botanic Gardens, Kew. 2020. State of the World's Plants and Fungi 2020.

Roycroft E, MacDonald AJ, Moritz C, Moussalli A, Portela Miguez R, Rowe KC. 2021. Museum genomics reveals the rapid decline and extinction of Australian rodents since European settlement. Proceedings of the National Academy of Sciences 118: e2021390118.

Scheele BC, Heard GW, Cardillo M, Duncan RP, Gillespie GR, Hoskin CJ, Mahony M, Newell D, Rowley JJL, Sopniewski J. 2023. An invasive pathogen drives directional niche contractions in amphibians. Nature Ecology & Evolution 7: 1682–1692.

Schulman L, Lahti K, Piirainen E, Heikkinen M, Raitio O, Juslén A. 2021. The Finnish Biodiversity Information Facility as a best-practice model for biodiversity data infrastructures. Scientific Data 8: 137.

Staude IR, Navarro LM, Pereira HM. 2020. Range size predicts the risk of local extinction from habitat loss. Global Ecology and Biogeography 29: 16–25.

Sterner BW, Gilbert EE, Franz NM. 2020. Decentralized but Globally Coordinated Biodiversity Data. Frontiers in Big Data 3: 519133.

Tandon A. 2021. Analysis: The lack of diversity in climate-science research. Carbon Brief. (21 December 2022; https://www.carbonbrief.org/analysis-the-lack-of-diversity-in-climate-science-research/).

Tauch A, Al-Dilaimi A. 2019. Bioinformatics in Germany: toward a national-level infrastructure. Briefings in Bioinformatics 20: 370–374.

Thelwall M, Maflahi N. 2022. Research coauthorship 1900–2020: Continuous, universal, and ongoing expansion. Quantitative Science Studies 3: 331–344.

Tydecks L, Jeschke JM, Wolf M, Singer G, Tockner K. 2018. Spatial and topical imbalances in biodiversity research. PloS One 13: e0199327.

Urban MC, Travis JMJ, Zurell D, Thompson PL, Synes NW, Scarpa A, Peres-Neto PR, Malchow A-K, James PMA, Gravel D, De Meester L, Brown C, Bocedi G, Albert CH, Gonzalez A, Hendry AP. 2022. Coding for Life: Designing a Platform for Projecting and Protecting Global Biodiversity. BioScience 72: 91–104.

Vattakaven T, George RM, Balasubramanian D, Réjou-Méchain M, Muthusankar G, Ramesh BR, Prabhakar R. 2016. India Biodiversity Portal: An integrated, interactive and participatory biodiversity informatics platform. Biodiversity Data Journal e10279.

Warren R, Price J, Graham E, Forstenhaeusler N, VanDerWal J. 2018. The projected effect on insects, vertebrates, and plants of limiting global warming to 1.5°C rather than 2°C. Science 360: 791–795.

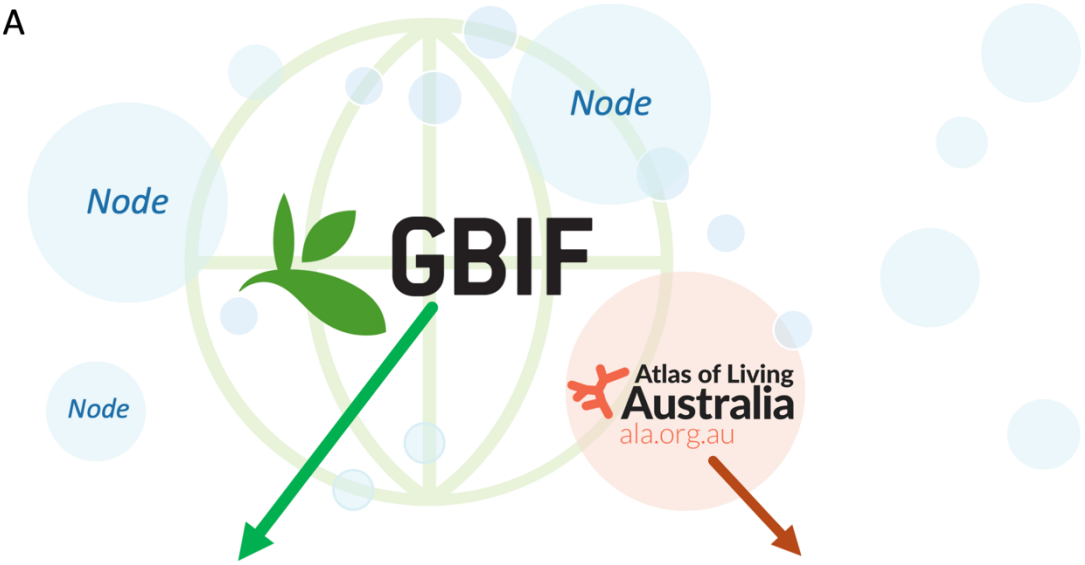
Westgate M, Stevenson M, Kellie D, Newman P. 2023. galah: Atlas of Living Australia (ALA) Data and Resources in R. R Package Version 1.4.0. https://CRAN.R-project.org/package=galah.

Wieczorek J, Bloom D, Guralnick R, Blum S, Döring M, Giovanni R, Robertson T, Vieglais D. 2012. Darwin Core: an evolving community-developed biodiversity data standard. PloS One 7: e29715.

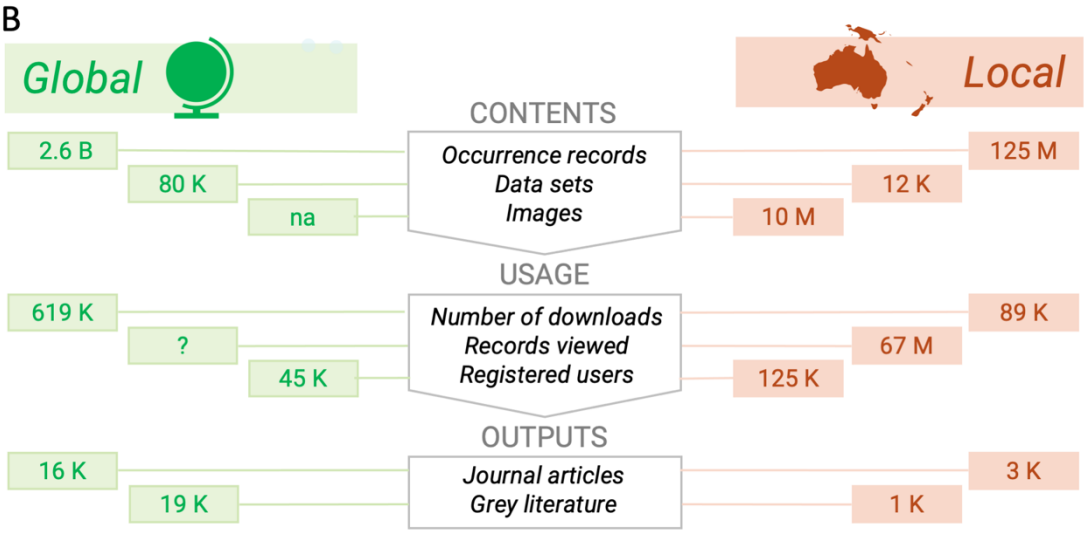
Figures

[Figures are also uploaded separately as pdf files]

A



B



C

Region	Atlas Name (link)	Records
Global	Global Biodiversity Information Facility (gbif.org)	1,817 M
Australia	Atlas of Living Australia (www.ala.org.au)	114 M
Austria	Biodiversitäts-Atlas Österreich (biodiversityatlas.at)	8 M
Brazil	Sistemas de Informações Sobre a Biodiversidade Brasileira (sibbr.gov.br)	24 M
Estonia	eElurikkus (elurikkus.ee)	7 M
France	Portail Français D'accès Aux Données D'observation Sur Les Espèces (openobs.mnhn.fr)	128 M
Guatemala	Sistema Nacional de Información sobre Diversidad Biológica de Guatemala (snib.conap.gob.gt)	4 M
Portugal	GBIF Portugal (www.gbif.pt)	16 M
Spain	GBIF Spain (www.gbif.es)	38 M
Sweden	Swedish Biodiversity Data Infrastructure (biodiversitydata.se)	103 M
UK	National Biodiversity Network (nbn.org.uk)	208 M

Figure 1: *GBIF (Global Biodiversity Information Facility) and its nodes, including Living Atlases. A – conceptual representation of the relationship between GBIF and its nodes, with*

the Atlas of Living Australia (ALA) as an example of a local node. B – comparison of the Global Biodiversity Information facility (GBIF) and the Atlas of Living Australia (ALA). Numbers represent rounded total counts for 2023. Data from: https://analytics-files.gbif.org/2024-01-01/download/csv/ (download_year.csv), https://www.gbif.org/resource/search?contentType=literature, https://doi.org/10.54102/ala.87540. C – total counts of records of GBIF and 10 Living Atlases accessible via the galah R package (https://galah.ala.org.au; accessed March 2023; (Westgate et al. 2023)). Discrepancies between numbers reported in panels B and C are due to differences in record summarisation methods between platforms.

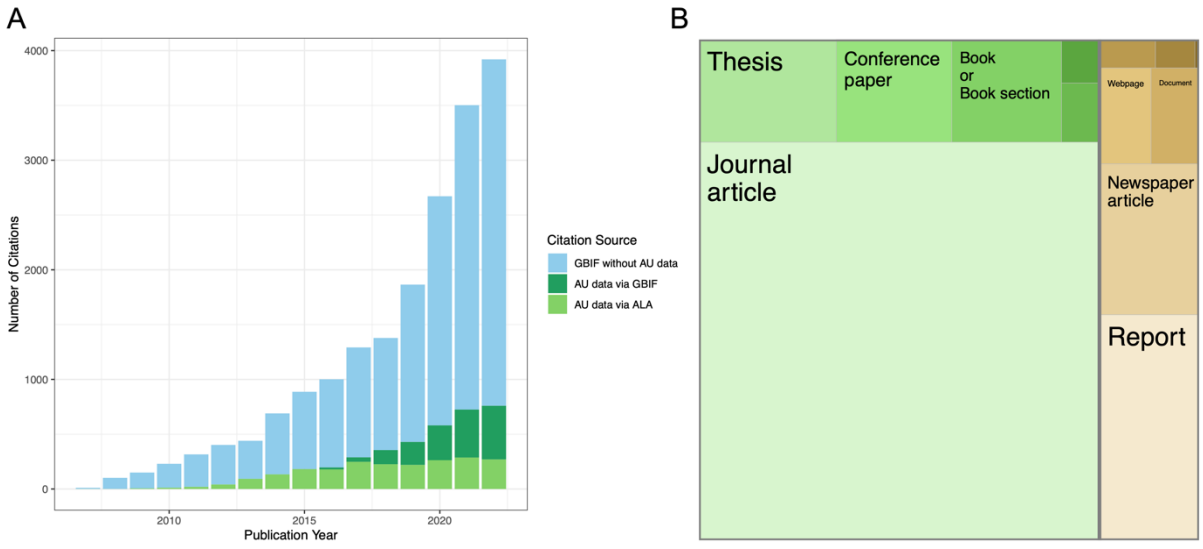


Figure 2: *Overview of the Atlas of Living Australia (ALA)-related outputs – historical trends and output types. A – Annual citations of the GBIF curated outputs with and without ALA data in relation to trends in citations of Australian data via ALA only. B – Composition of the*

ALA-related outputs citing, using or mentioning the ALA up to 2023, based on the ALA-curated records.

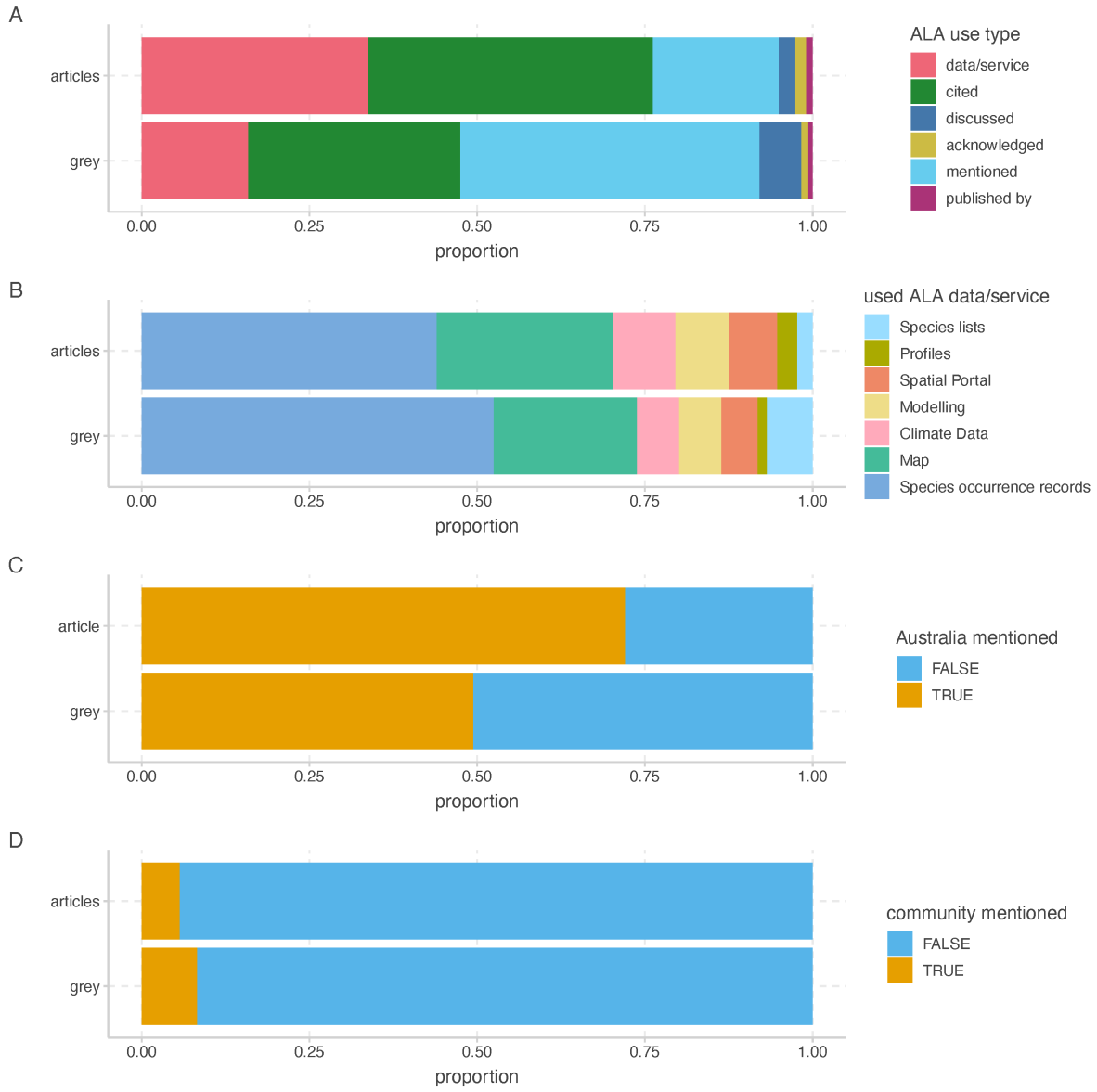


Figure 3: *Comparisons of the scope of the Atlas of Living Australia (ALA)-related outputs for outputs classified as journal articles (article) and other outputs (grey). A. – The role of ALA in the outcomes as coded by manually assigned tags. B. – Types of directly used ALA resources or services. C. – Australian focus as indicated by the use of Australia-specific words in the*

titles or abstracts. D. – Community and citizen science focus as indicated by the use of related words in the titles or abstracts.

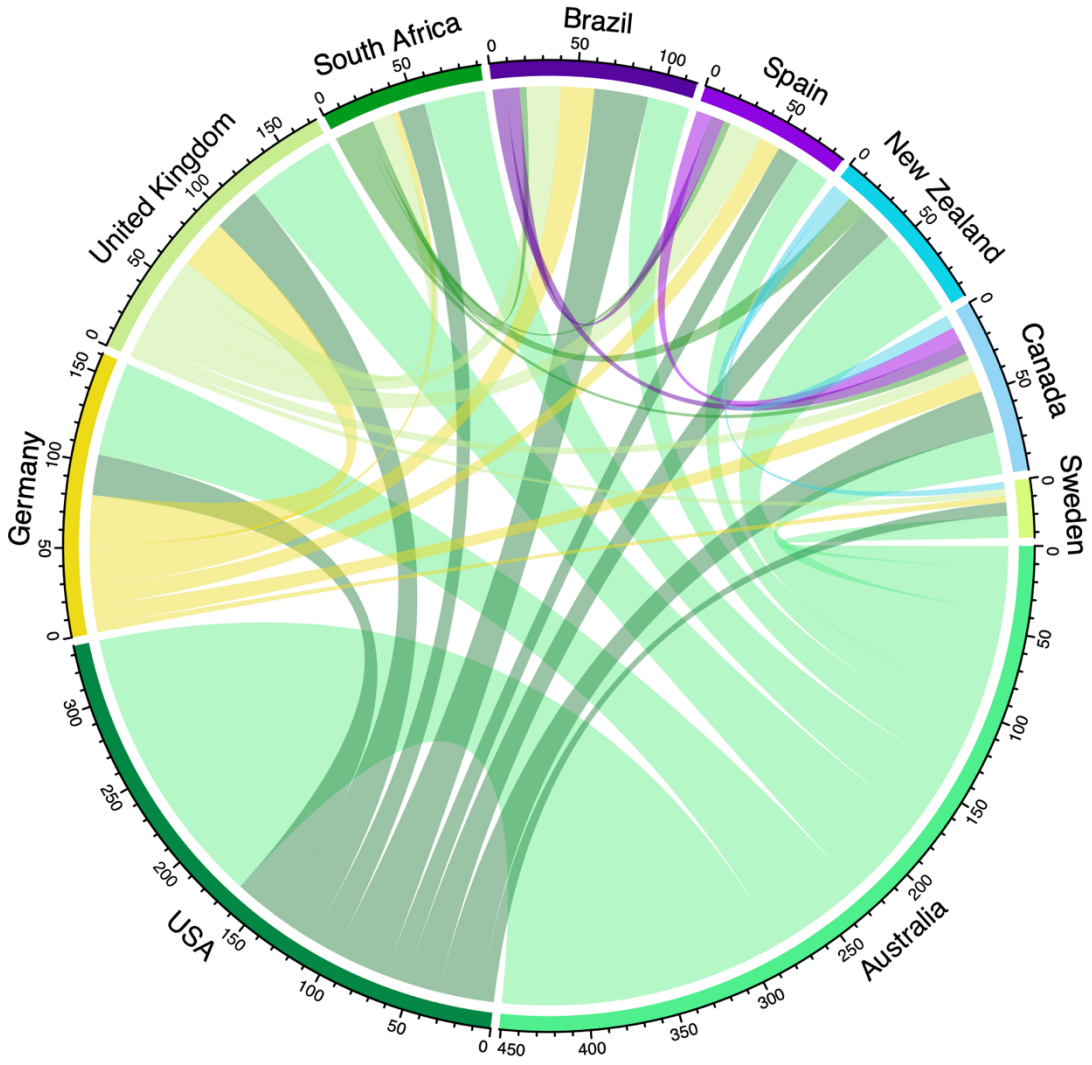


Figure 4: *Global authorship and collaboration. Collaboration links between the 10 most authorship-intensive countries. The links represent the number of co-authored journal articles and are coloured according to the dominant country; collaborations within countries are not shown.*

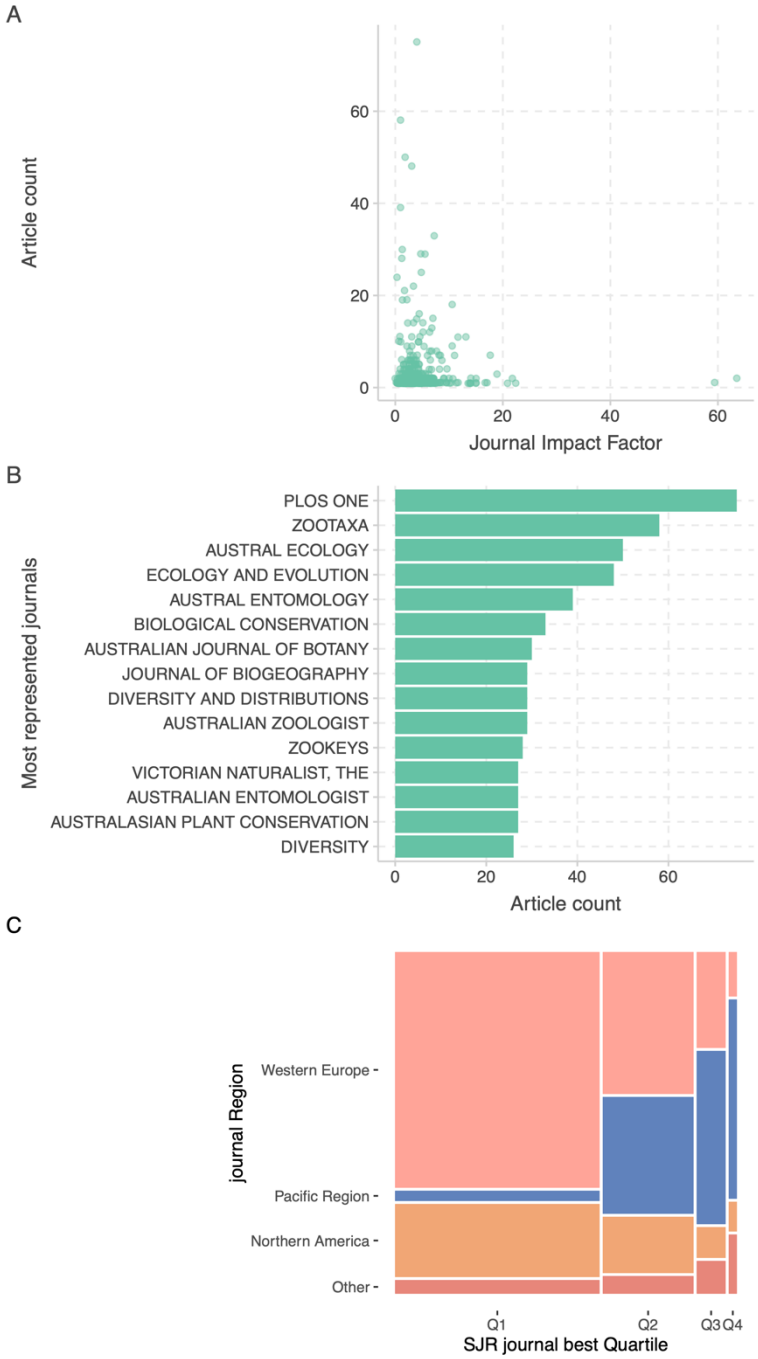


Figure 5: *Characteristics of journals of in which ALA-related journal articles were published. A. – Counts of articles in relation to journals Impact Factors (IF). B. – top 15 most frequent journals and archiving platforms. C. – counts of articles by journal SJR journal Region and Best Quartile.*

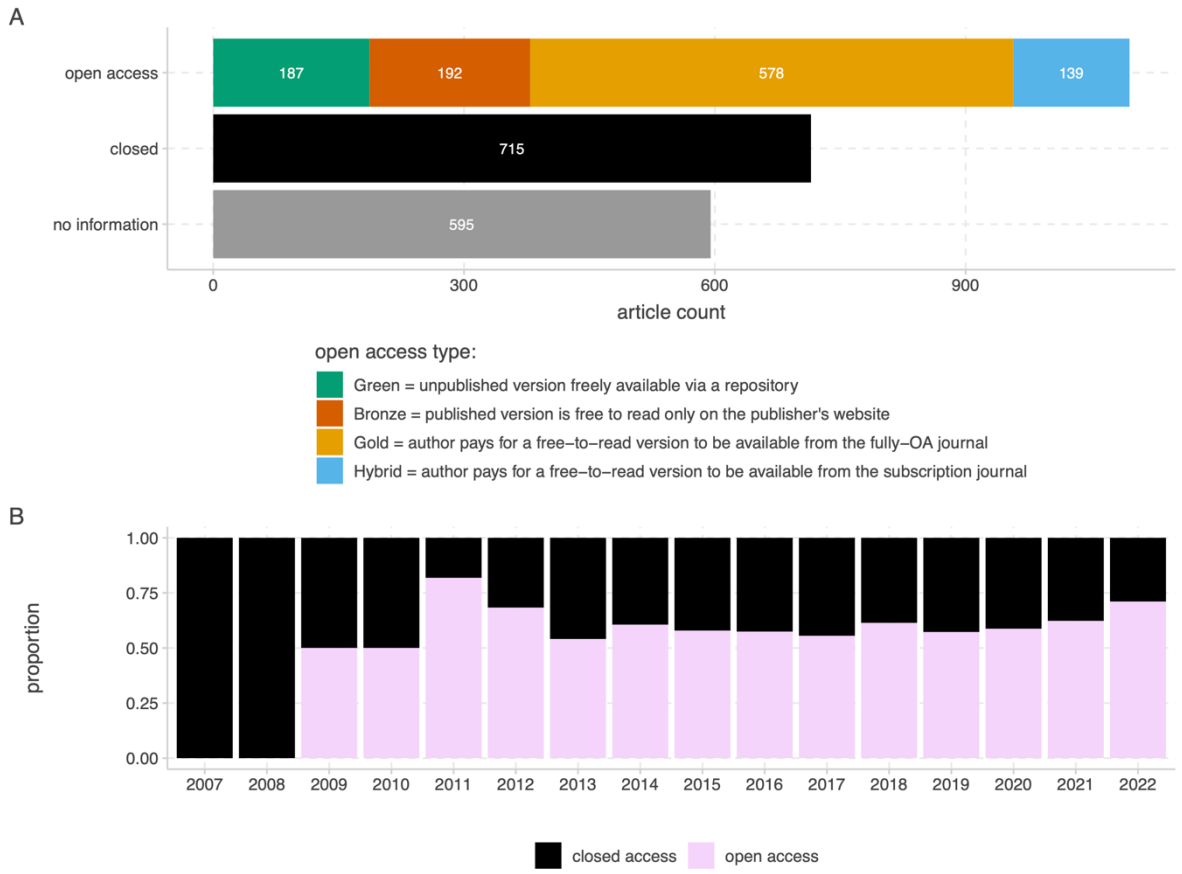


Figure 6: *Accessibility of the full-texts of ALA-related journal articles. A. – Counts of journal articles that have open access are published under a range of publishing options (open access type), as indicated by bar colours. "no information" indicates that no record exists for this subset of articles in the Unpaywall Open Access database. White numbers show counts of articles in each sub-category of open access. B. – Historical changes in the proportion of openly accessible journal articles (articles with no information on their accessibility status are not shown, all open access types are pooled).*

Going Global by Going Local: Impacts and Opportunities of Geographically Focused Data Integration

Supplementary Material 1

Supplementary Methods

This document section includes additional details of the methods presented in the main text, as well as supplementary Figure S1 and Figure S2 presenting results of network analyses included in the Supplementary File (Supplementary Material 2 – data and code alo available on GitHub).

ALA publication tracking methods

The Atlas of living Australia (ALA) has tracked articles that use, cite, or mention the ALA since 28 February 2018. The ALA stores these articles in a Zotero library along with article metadata. Articles range in publication date from 2007 – 2023.

To track relevant articles, the ALA has set up email notifications from literature searching databases. The ALA receives notifications from Google Scholar, ProQuest, Scopus, Isentia and Web of Science (Core Collection and Data Cite Index). The ALA also receives notifications about datasets and figures from CSIRO Mediaportal, Zenodo and Figshare. Email notifications are sent to a specific ALA reference email address.

The ALA tracks when their name, website or DOI are used in an article. The ALA is notified when the following key words are in an article:

- Atlas of Living Australia
- Atlas Living Australia
- ala.org.au
- 10.26197
- grid.506668.b
- https://ror.org/018n2ja79 (added 25/01/2022)

Data downloads made using the ALA website or the R package {galah} are minted a DOI beginning with 10.25197, and the suggested citation for the ALA is using grid.506668.b (before 25/01/2022) or https://ror.org/018n2ja79 (after 25/01/2022).
Articles are added to the Zotero library weekly. The publication tracking workflow was first established by Corinna Paeper and Ely Wallis in 2018. Publication tracking is currently maintained by Olivia Torresan, Dax Kellie, and Martin Westgate.

Manual record tagging

When an email is received, articles were manually checked to determine how the ALA was used or discussed in the article. If a manual check can verify that the ALA was cited or mentioned in the article, the article is added to Zotero with associated metadata and the full article PDF (if available).
One or more tags were manually assigned in Zotero to each article to indicate the ALA's relevance within the article. These tags were:
Tag description:

- "1 – ALA used" - Makes use of data in a quantitative analysis (e.g., ecological niche modelling, species distribution modelling). This can include use as supplemental data in a main or supplementary analysis
- "2 – ALA cited" - Cites a quantitative or qualitative fact derived from data (e.g. a statement or map of a given species' distribution)
- "3 – ALA discussed" - Discusses ALA as an infrastructure or the use of data
- "4 – ALA acknowledged" - Acknowledges the ALA but doesn't use or cite data (e.g., statements like "infrastructures like the ALA")
- "5 – ALA mentioned" - Unspecifically mentions ALA or the ALA portal
- "6 – ALA published" - Describes or talks about data published to the ALA.

Additional tags were also assigned to "1 – ALA Used" articles if it can be discerned which type of ALA data or ALA tools were used to retrieve the data (e.g., Australian Virtual Herbarium, Spatial Portal, Species occurrence records). Whether these tags were assigned varies by article, and there is limited documentation of how consistently they were assigned since tracking started in 2018.

Analyses software

For all analyses, we used R version 4.2.3 (2023-03-15) (R Core Team 2013) in RStudio (2023.03.0 Build 386; "Cherry Blossom" Release (3c53477a, 2023-03-09) for macOS) (RStudio Team 2022). R packages used in data processing and plotting include: tidyverse 2.0.0 (Wickham et al. 2019), here v.1.0.1 (Müller 2020), readr v.2.1.2 (Wickham et al. 2023), rvest v.1.0.2 (Wickham 2022a), httr v.1.4.4 (Wickham 2023), pdftools v.3.3.2 (Ooms 2023), stringr v.1.4.0 (Wickham 2022b), tidystringdist v.0.1.4 (Fay 2019), galah v.1.5.2 (Westgate et al. 2023), roadoi v.0.7.2 (Jahn 2022), rAltmetric v.0.7.0 (Ram 2012), openalexR v.1.0.2.9 (Aria and Le 2023), igraph v.1.3.1 (Csardi and Nepusz 2006), bibliometrix v.4.0.0 (Aria and Cuccurullo 2017), circlize v.0.4.15 (Gu et al. 2014), maps v.3.4.0 (Becker et al. 2022), ggplot2 v.3.3.6 (Wickham 2016), ggrepel v.0.9.1 (Slowikowski 2023), RColorBrewer v.1.1-3 (Neuwirth 2022) , ggraph v.2.0.5 (Pedersen 2022a), ggmosaic v.0.3.4 (Jeppson et al. 2023), ggthemes v.4.2.4 (Arnold 2021), ggthemr v.1.1.0 (Tobin 2020), patchwork v.1.1.1 (Pedersen 2022b).

Comparing the scope of journal articles and grey outputs

We first summarised the bibliographic records from Zotero database in terms of the number of records overall and across the years. We aggregated sixteen values of literature type into two main outcome types: journal articles (article) and grey outputs (grey, i.e., any outputs that are not journal articles). We then used the two outcome types for the comparisons of the literature properties that are applicable to both types (e.g., scope and authorship, see below) and to create an article-only data subset for additional article-specific analyses.
We then summarised the scope of the articles and grey outputs in terms how they use or credit ALA, using manually assigned tags described above (6 use-type tags: "1 - ALA used", "2 - ALA cited", "3 - ALA discussed" , "4 - ALA acknowledged", "5 - ALA mentioned", "6 - ALA published"; resource-type tags: “Species occurrence records", "Map" , "Modelling", "Climate Data", "Spatial Portal", "Species lists", "Profiles"). We discarded remaining tags which were context-specific and sparse.
We also searched titles and abstracts for words that indicate that they are focused on Australian species, ecosystems or locations, using str_detect function from the stringr

package with pattern = "austral|tasman|victoria|queensland|northern territiory|new south wales |nsw|vic|qld|sunshine coast|canberra|mudgee|pilbara|illawarra|sydney|Melbourne |perth|adelaide|brisbane|great barrier reef|snowy mountains|murray-darling|gippsland |aboriginal|kangaroo|koala|platypus|echidna|quoll|eucalypt". Records with positive matches were counted as potentially being relevant to Australia. We also searched titles and abstracts for words that indicate that broader community was involved in the work, for example as citizen science volunteers or co-researchers, with pattern = "citizen|volunteer|crowdsourc" ("community" was not used as a keyword as it would mostly retrieve works related to biological communities of species other than humans).

Authorship diversity and collaboration

We processed authors names stored in Zotero bibliographic network to initialise all first names, because majority of the names were stored in this format. We then used authorship information to calculate the total number of authors overall, total number of unique author names. We compared the sizes of the authorship teams (number of authors per output) between literature and across the years and noted outputs with unusually big team sizes (>50). We constructed the co-authorship network among the ALA-related outputs using the igraph package. We compared connectedness of authorship networks for articles and grey outputs by calculating indexes of node centrality, graph density, transitivity (clustering coefficient) and modularity (relative cluster density).

Affiliated country of journal article authors

Most of the grey output bibliographic records and many of the journal articles records did not contain author affiliation data or the data was inconsistent. For the articles with DOI, we retrieved affiliation information from OpenAlex platform API (https://docs.openalex.org/api) using openalexR package. We courted number of unique author affiliation countries in the dataset and across individual articles to measure

internationality. We noted the most frequently appearing countries in the authorship of journal articles. We looked at the patterns of collaborations by creating the matrix of co-authorship links among all countries using bibliometrix package. We then created a chord diagram of collaborations among the top 10 dominant countries using circlize package.

Publication patterns of journal articles

We retrieved information on Journal Impact Factors (IF) from a "JCR-2021-Impact-Factor-PDF.pdf" file available from Journal Citation Reports (JCR) at https://impactfactorforjournal.com/wp-content/uploads/2022/06/JCR-2021-Impact-Factor-PDF.pdf. We extracted text from the pdf file using pdftools package and converted the context into a table of 10,316 journal names and IF values. We then used capitalised journal names to merge IF values with our main bibliographic data frame of ALA-related outputs. We then summarised the distribution of IF in the dataset and compared the list of publication journals with IF > 15 with the list of the top 15 most frequent journals and publishing platforms. We retrieved information on journal ranking by downloading journal data from SCImago Journal & Country Rank (SJR) website at https://www.scimagojr.com/journalrank.php. The resulting .csv file contained journal rank, quantile information, region, and subject categories, among other data. We used capitalised journal names to merge this information with our main bibliographic data frame. We summarised journal best quartile and journal regions for our dataset as indicators of the impact and scope of ALA-related outputs, respectively.

Accessibility of journal articles

To retrieve the data on the open access status of individual articles, we accessed Unpaywall database API (https://unpaywall.org/products/api) using the roadoi package. We fetched and merged the data using article DOI as the unique and universal output identifier. We summarised the frequencies of the Gold, Bronze, Hybrid, Green, and closed publishing

article status and visualised historical changes in the proportion of any open versus closed outputs across the years.

Impact of journal articles

We used the rcrossref package to get citation counts from the Crossref OpenURL API (https://www.crossref.org/documentation/retrieve-metadata/openurl/). We fetched and merged the data using article DOI as the unique and universal output identifier and summarised the distribution of citations across the years. To retrieve altmetric data (attention outside traditional academic publications), we customised functions from an archived version of rAltmetric package (this package is no longer maintained or available from CRAN). We then used the article DOI for fetching data from the Altmetric platform via API (https://api.altmetric.com/). Retrieved data included weighted overall Altmetric Score (Altmetric Attention Score) and counts of sources that have mentioned the publication by source category (e.g., policies, patents, news, blogs, Twitter accounts, Facebook pages, Wikipedia citations, and many other). We summarised the distribution of Altmetric scores across the years using unadjusted Altmetric scores and scores adjusted for journal context (comparing to other articles from the same journals). We also identified articles that were cited in policies, patents, Wikipedia, and news outlets, or that received outstanding social media attention. For more details, see code descriptions in the Supplementary File.

Supplementary References

Aria M, Cuccurullo C. 2017. bibliometrix: An R-tool for comprehensive science mapping analysis. Journal of Informetrics.

Aria M, Le T. 2023. openalexR: Getting Bibliographic Records from 'OpenAlex' Database Using 'DSL' API.

Arnold JB. 2021. ggthemes: Extra Themes, Scales and Geoms for 'ggplot2'.

Becker OS code by RA, Minka ARWR version by RBE by TP, Deckmyn A. 2022. maps: Draw Geographical Maps.

Csardi G, Nepusz T. 2006. The igraph software package for complex network research. InterJournal Complex Systems: 1695.

Fay C. 2019. tidystringdist: String Distance Calculation with Tidy Data Principles.

Gu Z, Gu L, Eils R, Schlesner M, Brors B. 2014. circlize implements and enhances circular visualization in R. Bioinformatics 30: 2811–2812.

Jahn N. 2022. roadoi: Find Free Versions of Scholarly Publications via Unpaywall.

Jeppson H, Hofmann H, Cook D. 2023. ggmosaic: Mosaic Plots in the 'ggplot2' Framework.

Müller K. 2020. here: A Simpler Way to Find Your Files.

Neuwirth E. 2022. RColorBrewer: ColorBrewer Palettes.

Ooms J. 2023. pdftools: Text Extraction, Rendering and Converting of PDF Documents.

Pedersen TL. 2022a. ggraph: An Implementation of Grammar of Graphics for Graphs and Networks.

Pedersen TL. 2022b. patchwork: The Composer of Plots.

R Core Team. 2013. R: A language and environment for statistical computing.

Ram K. 2012. rAltmetric: Retrieves altmerics data for any published paper from altmetrics.com.

RStudio Team. 2022. RStudio: Integrated Development for R.

Slowikowski K. 2023. ggrepel: Automatically Position Non-Overlapping Text Labels with 'ggplot2'.

Tobin C. 2020. ggthemr: Themes for 'ggplot2'.

Westgate M, Stevenson M, Kellie D, Newman P. 2023. galah: Atlas of Living Australia (ALA) Data and Resources in R.

Wickham H. 2016. ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag New York.

Wickham H. 2022a. rvest: Easily Harvest (Scrape) Web Pages.

Wickham H. 2022b. stringr: Simple, Consistent Wrappers for Common String Operations.

Wickham H. 2023. httr: Tools for Working with URLs and HTTP.

Wickham H, Averick M, Bryan J, Chang W, McGowan LD, François R, Grolemund G, Hayes A, Henry L, Hester J, Kuhn M, Pedersen TL, Miller E, Bache SM, Müller K, Ooms J, Robinson D, Seidel DP, Spinu V, Takahashi K, Vaughan D, Wilke C, Woo K, Yutani H. 2019. Welcome to the tidyverse. Journal of Open Source Software 4: 1686.

Wickham H, Hester J, Bryan J. 2023. readr: Read Rectangular Text Data.

Supplementary Figures

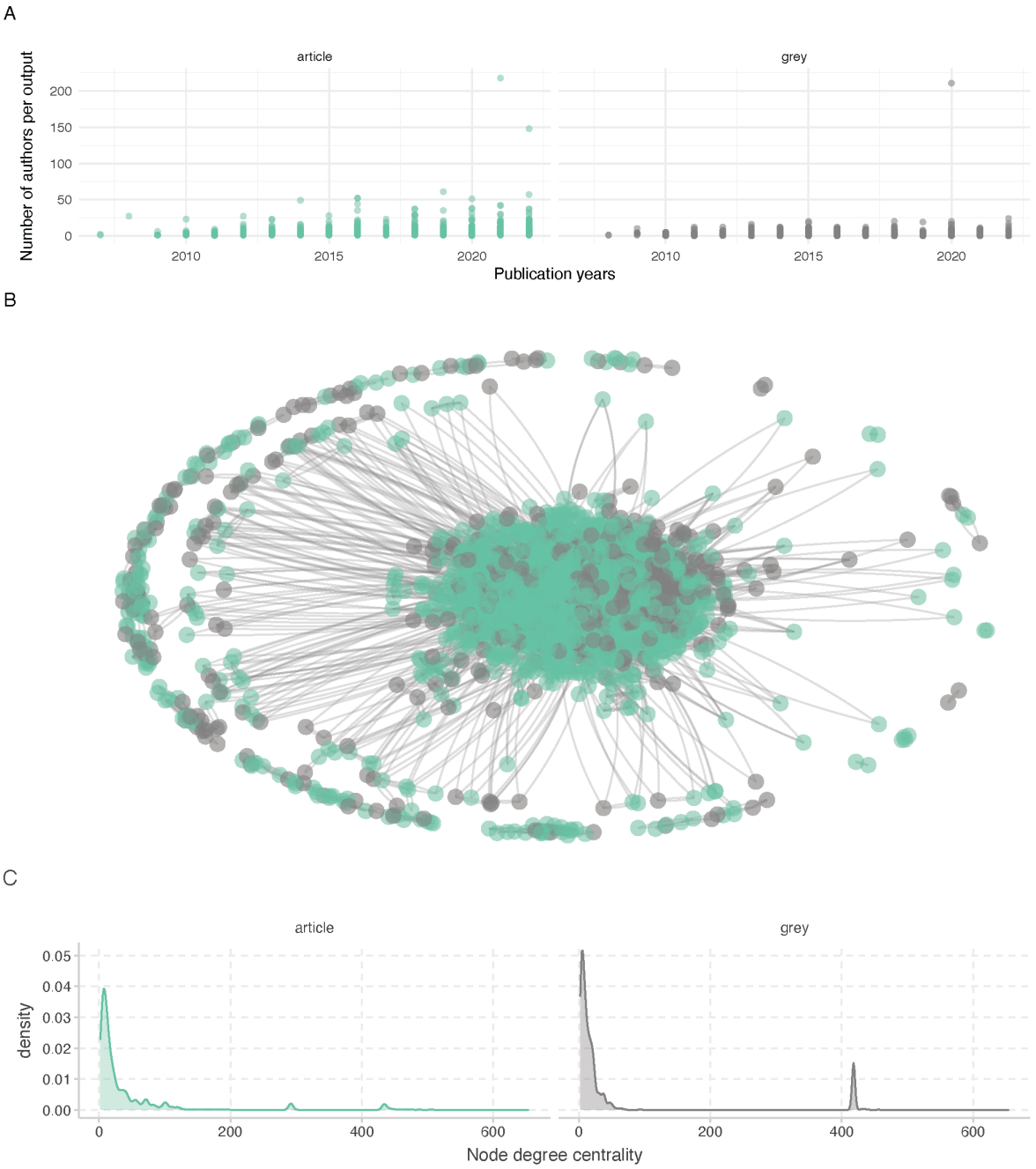


Figure S1: *Authorship and collaboration for ALA-related outputs. A. – Historical trends in author team sizes for journal articles and other output types (*grey outputs typically have smaller author team sizes: difference estimate = -0.541, SE = 0.019, *z* = -27.98, p < 0.001; slope estimate for journal articles = 0.7208, SE = 0.148, *t* = 4.85, *df* = 2,401, p < 0.001; slope estimate for grey outputs = 0.141, SE = 0.190, *t* = 0.742, *df* = 1210, p = 0.458). *B. – Co-authorship network of all outputs as nodes (sea green circles – journal articles, grey circles – other output types) and connections (lines) representing shared authors. C. – Density*

distributions of individual nodes degrees of the centrality of co-authorship networks for journal articles and other output types.

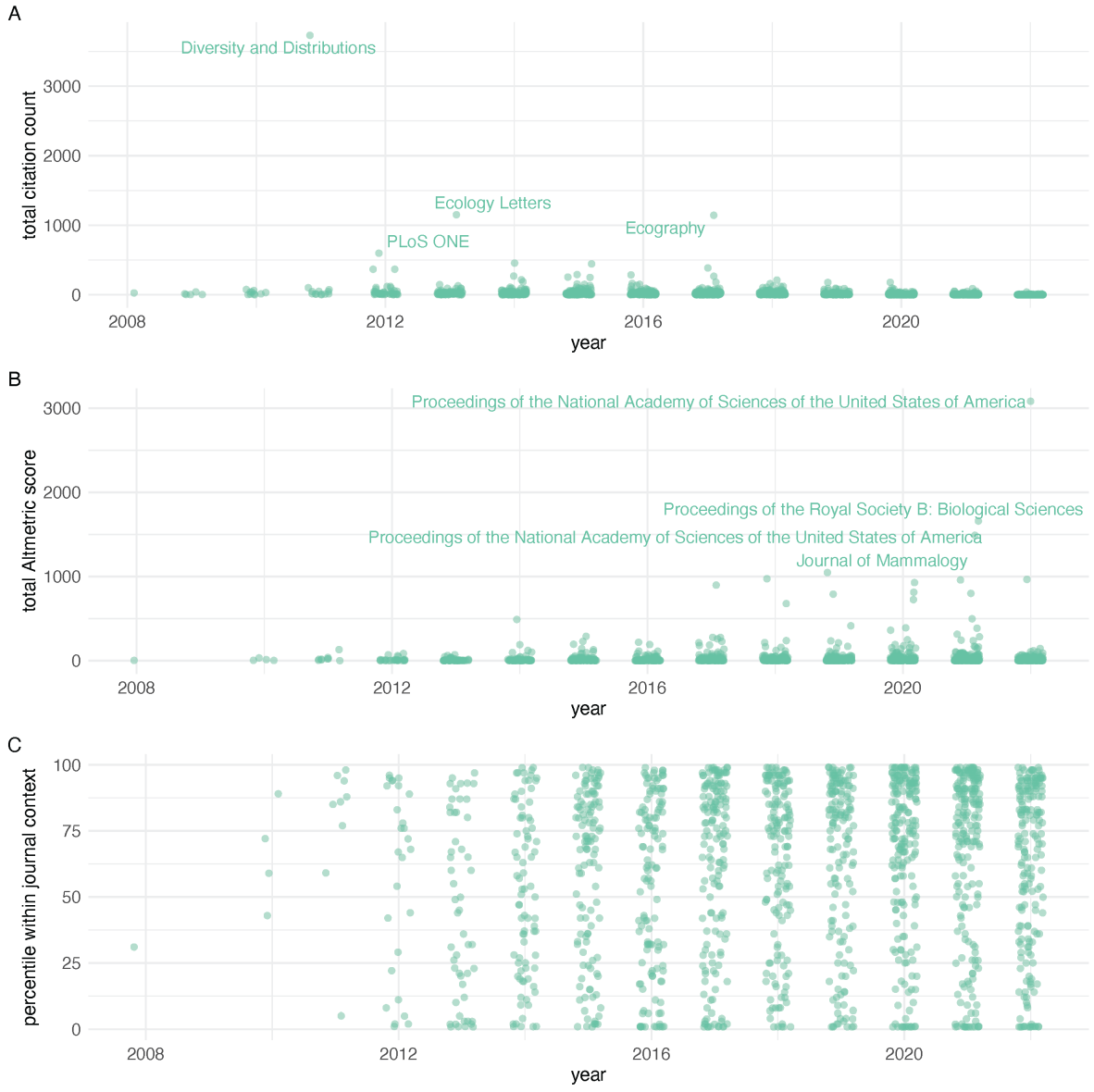


Figure S2: *Indicators of academic and non-academic impacts of the Atlas of Living Australia (ALA)-related journal articles. A. – Counts of citations by academic literature by journal article publication year. Journal names are shown as labels for top-performing articles. B. – total composite Altmetric scores by journal article publication year. Journal names are shown as labels for top-performing articles. C. – total composite Altmetric scores presented as journal percentile by journal article publication year. Data points before 2008 and after 2022 are trimmed off.*

Going Global by Going Local: Impacts and Opportunities of Geographically Focused Data Integration

Supplementary Material - data and code

Code

AUTHOR
Malgorzata Lagisz

PUBLISHED
March 29, 2024

Analyses details

This document includes complete processing code for analyses presented in the main text, as well as additional details and results. This document is also intended to be rendered into an html file, but for logistical reasons not all analyses results are shown in that document (especially collaboration networks which take long time to run and plot and thus are not automatically evaluated. Instead, relevant subsequent code chunks can be run by loading already fetched and/or preprocessed data objects saved as intermediate files).

The main data source used in this project is an ALA-curated impact literature collection, hosted on Zotero and exported as .csv file on 2023/03/30 into data/2023-03-30_ALA_cited.csv. Additional information was added consecutively from several other sources (OpenAlex, Unpaywall, Crossref, Altmetrics, Scimago), as described in this document.

▼ Code

```
sessionInfo()
```

```
R version 4.3.2 (2023-10-31)
Platform: x86_64-apple-darwin20 (64-bit)
Running under: macOS Sonoma 14.4

Matrix products: default
BLAS:   /Library/Frameworks/R.framework/Versions/4.3-x86_64/Resources/lib/libRblas.0.dylib
LAPACK: /Library/Frameworks/R.framework/Versions/4.3-x86_64/Resources/lib/libRlapack.dylib;  LA

locale:
[1] en_US.UTF-8/en_US.UTF-8/en_US.UTF-8/C/en_US.UTF-8/en_US.UTF-8

time zone: Australia/Sydney
tzcode source: internal

attached base packages:
[1] stats     graphics  grDevices utils     datasets  methods   base

loaded via a namespace (and not attached):
 [1] htmlwidgets_1.6.4 compiler_4.3.2    fastmap_1.1.1     cli_3.6.2
 [5] tools_4.3.2       htmltools_0.5.7   rstudioapi_0.15.0 yaml_2.3.8
 [9] rmarkdown_2.25    knitr_1.45        jsonlite_1.8.8    xfun_0.42
[13] digest_0.6.34     rlang_1.1.3       evaluate_0.23
```

▼ Code

```
knitr::opts_chunk$set(echo = TRUE, message = FALSE, warning = FALSE)
#devtools::install_github("massimoaria/bibliometrix") # to install the bibliometrix most
recent version from GITHUB
#install.packages('rAltmetric') #devtools::install_github("ropensci/rAltmetric")

pacman::p_load(tidyverse,
               here,
               rgbif,
               visdat, #for vis_miss
               roadoi, #https://cran.r-project.org/web/packages/roadoi/vignettes/intro.html
               corpcor,
               naniar,
               rvest,
               httr,
               httr2,
               bibliometrix,
               rcrossref,
               networkD3,
               stringr,
               tibble,
               tidystringdist,
               synthesisr,
               snakecase,
               tibble,
               ggplot2,
               RColorBrewer,
               rAltmetric,
               rlist,
               maps,
               mapproj,
               purrr,
               janitor,
               ggthemr,
               igraph,
               circlize,
               patchwork,
               pdftools,
               treemapify,
               ggraph,
               ggridges,
               ggthemes)

#install.packages("remotes")

#remotes::install_github("massimoaria/openalexR")
#install.packages("openalexR")
library(openalexR) #not on CRAN

#install.packages("devtools")
#devtools::install_github("haleyjeppson/ggmosaic")
library(ggmosaic)

#devtools::install_github('Mikata-Project/ggthemr')
library(ggthemr)

ggthemr('pale') #select one ggplot theme to be used


#install.packages("galah")
#install.packages("remotes")
#remotes::install_github("AtlasOfLivingAustralia/galah")
library(galah) #for gathering record counts from Living Atlases
```

Load and clean initial dataset for ALA impact assessment

Load original Zotero reference collection from a csv file.

▼ Code

```
dat <- read.csv(here("data", "2023-03-30_ALA_cited.csv"))
#dim(dat) #3866 87
#names(dat)
#table(dat$Publication.Year, useNA = "always") #92 in 2023, 10 missing year (webpages)
dat <- dat %>% filter(Publication.Year != 2023) #remove 2023 and records with missing year
#dim(dat) #3764 87

#add a column in the main data frame article vs. grey
dat %>% mutate(article_grey = case_when(Item.Type == "journalArticle" ~ "article",
                                        Item.Type != "journalArticle" ~ "grey")) -> dat
```

Make simple tables summarizing key selected variables.

▼ Code

```
table(dat$article_grey) #2406 articles and 1358 grey literature
table(dat$article_grey, dat$Abstract.Note=="") # only 758 of grey have abstract, 2059
articles have abstract
table(dat$Item.Type, useNA = "always")
#hist(dat$Publication.Year) #make graph by year, below
table(dat$Language, useNA = "always") #mosly en_AU, needs cleaning, below
#table(dat$Place, useNA = "always") #mostly empty, super messy - not usable
#table(dat$Manual.Tags, useNA = "always") #often more than one tag (different types) per
item - split later
table(str_starts(dat$DOI, "10.")) #2139 are proper DOI numbers (starting with 10.), other
are "" or "-"
table(str_starts(dat$DOI, "10."), dat$article_grey) #2070 out of 2406 articles have proper
DOI numbers (starting with 10.)
```

Clean DOI data column.

▼ Code

```
dat$DOI[dat$DOI == "-"] <- "" #replace with empty value for consistency
dat$DOI[dat$DOI == "n/a"] <- "" #replace with empty value for consistency
dat$DOI <- gsub("https://doi.org/", "", dat$DOI, fixed = TRUE)  #replace with ""
dat$DOI <- gsub("http://dx.doi.org/", "", dat$DOI, fixed = TRUE)  #replace with ""
length(unique(dat$DOI)) #2171 unique values
#table(str_starts(dat$DOI, "10."), dat$Item.Type) #DOI by Item.Type
```

Clean Language data column.

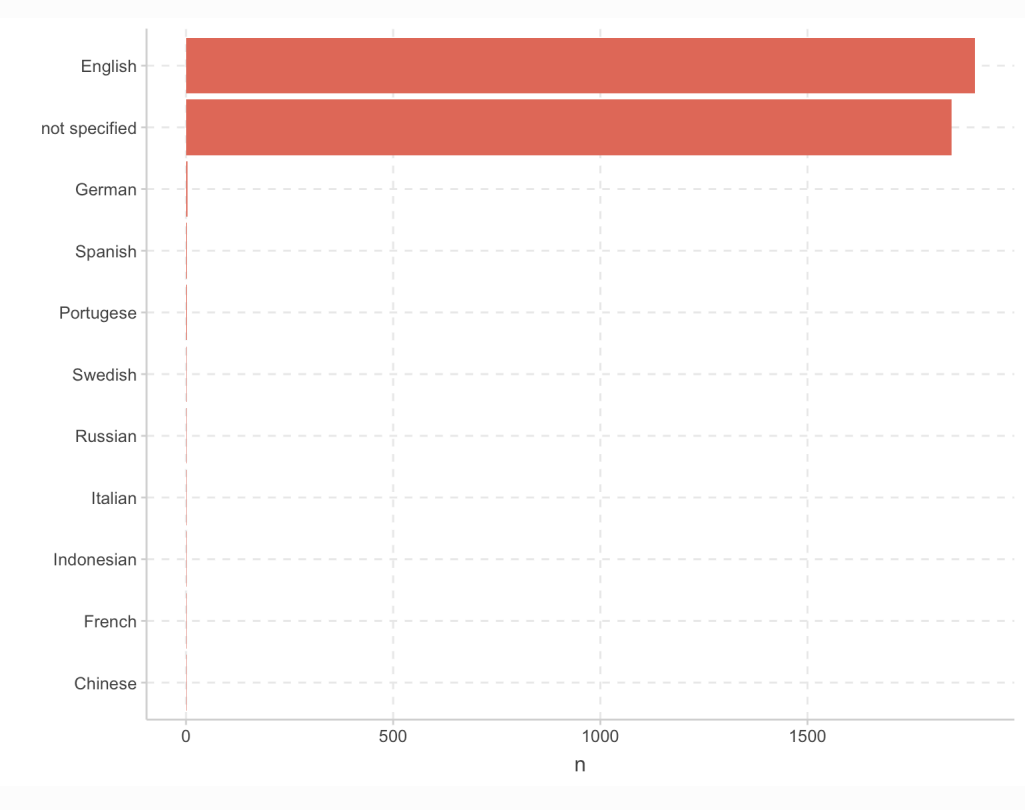
▼ Code

```
#table(dat$Language, useNA = "always") #needs cleaning, below
dat$Language <- replace(dat$Language, dat$Language %in% c("en","EN", "ENG", "en_AU",
"en_US", "en-AU", "en-GB", "en_GB", "en-IN", "en-us", "en-US", "eng", "dcterms.RFC4646; en-
AU"), "English")
dat$Language <- replace(dat$Language, dat$Language %in% c("de"), "German")
dat$Language <- replace(dat$Language, dat$Language %in% c("fr"), "French")
dat$Language <- replace(dat$Language, dat$Language %in% c("es"), "Spanish")
dat$Language <- replace(dat$Language, dat$Language %in% c("por"), "Portugese")
dat$Language <- replace(dat$Language, dat$Language %in% c("id"), "Indonesian")
dat$Language <- replace(dat$Language, dat$Language %in% c("IT"), "Italian")

#table(dat$Language, useNA = "always") #much better, but 1847 records have no language info
table(dat$Language)["English"] #1902 records with English
#sum(dat$Language=="", na.rm  = TRUE) #1847 records without language info (extra 12 are NA)
dat$Language <- replace(dat$Language, dat$Language %in% c(""), "not specified")
dat$Language <- replace(dat$Language, is.na(dat$Language), "not specified")
#table(dat$Language, useNA = "always") #much better, but 2003 records as "not specified"
```

Make a simple bar plot for languages.

▼ Code

```
dat %>%
  count(Language) -> count_Language
ggplot(count_Language, aes(x = reorder(Language, n), y = n)) + geom_bar(stat = "identity") +
  coord_flip() +
  xlab("")
```



Language: 1904 out of 3764 are in English, but for 1847 records the language info is missing.

Place information - 3049 recorded (messily, mostly for for grey literature), but majority is missing (715).

▼ Code

```
#table(dat$Place, useNA = "always") #needs cleaning, below
#table(dat$Place == "") #number of records that have recorded any info on the place - not
very useful column
all_countries <- str_c(unique(iso3166$sovereignty), collapse = "|") #use iso3166 data set
df <- dat #make a copy
#all_countries <- str_c(unique(world.cities$country.etc), collapse = "|")
df$country <- sapply(str_extract_all(df$Place, all_countries), toString)
table(df$country, useNA = "always") #needs much more cleaning
df$country <- replace(df$country, df$country %in% c("Australia, Australia"), "Australia")
#table(df$country=="", df$article_grey, useNA = "always") #place is recorded mostly for 509
grey lit (847 missing), 0 for articles
```

Where Country is recorded, 374 are from Australia.

Simple bar plot for places recorded for grey literature.

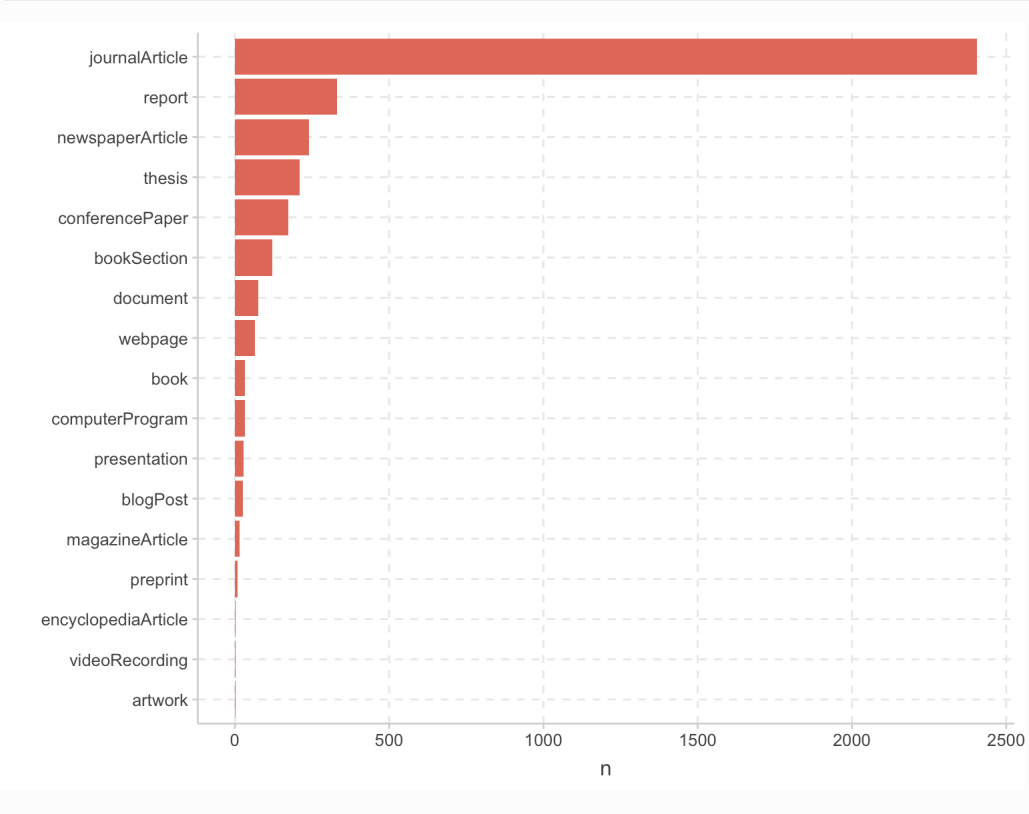
▼ Code

```
df %>%
  filter(article_grey == "grey") %>%
  count(country) -> count_country
ggplot(count_country, aes(x = reorder(country, n), y = n)) + geom_bar(stat = "identity") +
  coord_flip() +
  xlab("")
```

Simple bar plot for literature type.

▼ Code

```
dat %>%
  count(Item.Type) -> count_lit_type
ggplot(count_lit_type, aes(x = reorder(Item.Type, n), y = n)) + geom_bar(stat = "identity")
+
  coord_flip() +
  xlab("")
```



Clean string characters - standardise common foreign and formatting characters in the Author, Title and Abstract.Note columns.

▼ Code

```
#only use latin-ascii characters
dat$Author <- stringi::stri_trans_general(dat$Author, 'latin-ascii')
dat$Title <- stringi::stri_trans_general(dat$Title, 'latin-ascii')
dat$Abstract.Note <- stringi::stri_trans_general(dat$Abstract.Note, 'latin-ascii')
#remove some leftover html formatting
dat$Title <- gsub("<i>", "", dat$Title)
dat$Title <- gsub("</i>", "", dat$Title)
```

Clean manual tags (Manual.Tags column contains various manually assigned tags to identify how ALA was used and what was done).

▼ Code

```
mtags <- str_split(dat$Manual.Tags, "; ")
#table(unlist(mtags))
#keep only the 6 tags starting with numerics (main use)
mtags_num <- unlist(mtags)[str_detect(unlist(mtags), "^[0-9]")]
count_mtags_num <- as.data.frame(table(mtags_num))
count_mtags_num
```

```
              mtags_num Freq
1         1 - ALA used 1007
2        2 - ALA cited 1418
3    3 - ALA discussed  143
4 4 - ALA acknowledged   51
5     5 - ALA mentioned 1023
6     6 - ALA published   31
```

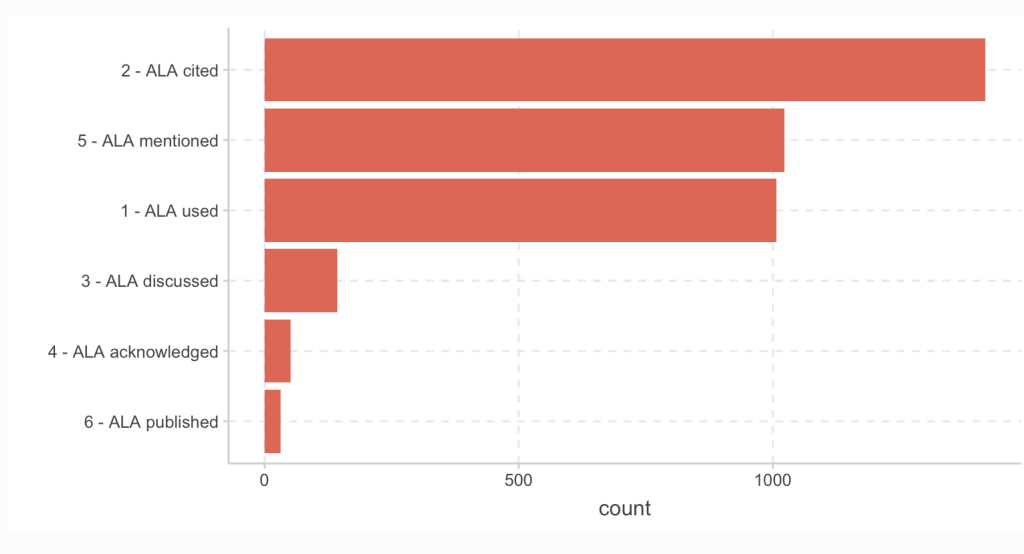
▼ Code

```
#str(count_mtags_num)
```

Plot frequencies of manual tags for ALA use type.

▼ Code

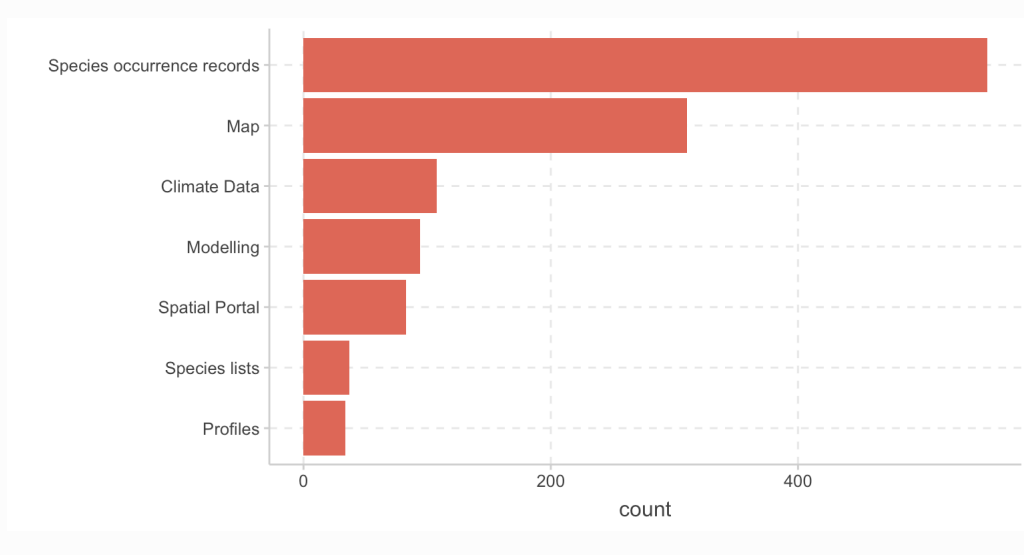
```
ggplot(count_mtags_num, aes(x = reorder(mtags_num, Freq), y = Freq)) + geom_bar(stat =
"identity") +
  coord_flip() +
  xlab("")  +
  ylab("count")
```



Plot manual tags counts for the categories of used ALA resources.

▼ Code

```
#Non-numeric manual tags assigned only to records where ALA was used ("1 - ALA used") - they
provide details of use
mtags_det <- unlist(mtags)[unlist(mtags) %in% c("Species occurrence records", "Map" ,
"Modelling", "Climate Data", "Spatial Portal", "Species lists", "Profiles")]
count_mtags_det <- as.data.frame(table(mtags_det))
#count_mtags_det

#plot
ggplot(count_mtags_det, aes(x = reorder(mtags_det, Freq), y = Freq)) + geom_bar(stat =
"identity") +
  coord_flip() +
  xlab("")  +
  ylab("count")
```



Australia-related content

Create columns indicating if Australia-related terms (e.g. Tasmania, Victoria, Queensland, NSW, Sydney) were mentioned in study title, abstract, or both.

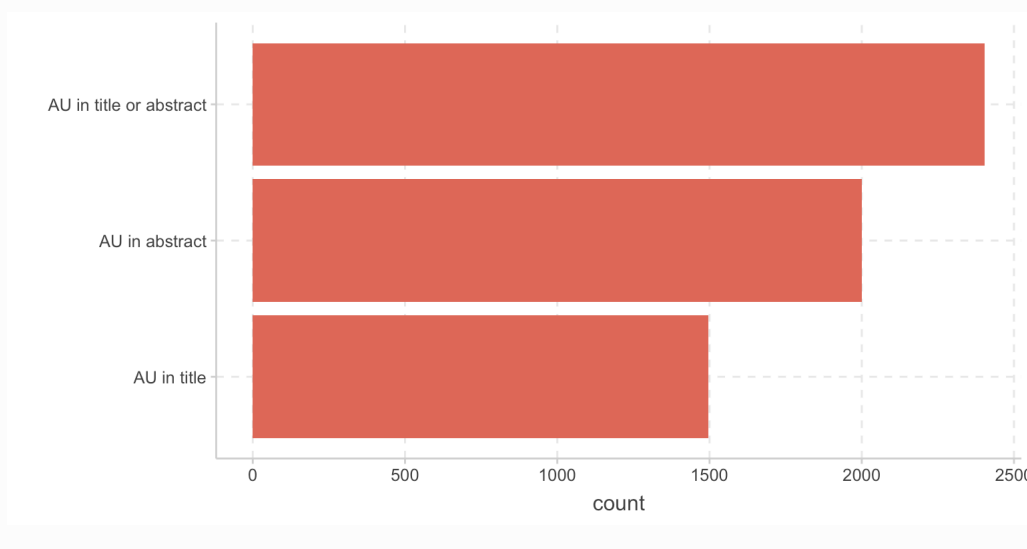
▼ Code

```
#create a new logical column with TRUE when AU-relevant terms mentioned in titles
dat$AU_title <- str_detect(string = str_to_lower(dat$Title), pattern =
"austral|tasman|victoria|queensland|northern territiory|new south wales|nsw|vic|qld|sunshine
coast|canberra|mudgee|pilbara|illawarra|sydney|melbourne|perth|adelaide|brisbane|great
barrier reef|snowy mountains|murray-
darling|gippsland|aboriginal|kangaroo|koala|platypus|echidna|quoll|eucalypt")
#dat$Title[dat$AU_title == FALSE] lists that that did not match
# table(dat$AU_title)

#create a new logical column with TRUE when AU-relevant terms mentioned in abstracts
dat$AU_abs <- str_detect(string = str_to_lower(dat$Abstract.Note), pattern =
"austral|tasman|victoria|queensland|northern territiory|new south wales|nsw|vic|qld|sunshine
coast|canberra|mudgee|pilbara|illawarra|sydney|melbourne|perth|adelaide|brisbane|great
barrier reef|snowy mountains|murray-
darling|gippsland|aboriginal|kangaroo|koala|platypus|echidna|quoll|eucalypt")
# table(dat$AU_abs)

#create a new logical column with TRUE when AU-relevant terms mentioned in title or abstract
dat$AU_title_abs <- ifelse(dat$AU_title == "TRUE" | dat$AU_abs == "TRUE", "TRUE", "FALSE")
# table(dat$AU_title_abs)
```

Plot where (column / reference field) Australian context is mentioned.

▼ Code

```
#summarise and merge for plotting for 3 new columns across all records
count_AU <- as.data.frame(rbind("AU in title" = table(dat$AU_title), "AU in abstract" =
table(dat$AU_abs), "AU in title or abstract" = table(dat$AU_title_abs)))
colnames(count_AU) <- c("not metioned", "mentioned")
count_AU$Australia <- rownames(count_AU)

#plot
ggplot(count_AU, aes(x = reorder(Australia, mentioned), y = mentioned)) + geom_bar(stat =
"identity") +
  coord_flip() +
  xlab("")  +
  ylab("count")
```



Citizen science focus - by literature type.

Create columns recording if citizen science (CS) or related terms (e.g. volunteer, crowdsourcing) were mentioned in study title, abstract, or both (cannot use "community" because it often means "ecological community" (of various species) not "human community" (general public or local residents, etc.)).

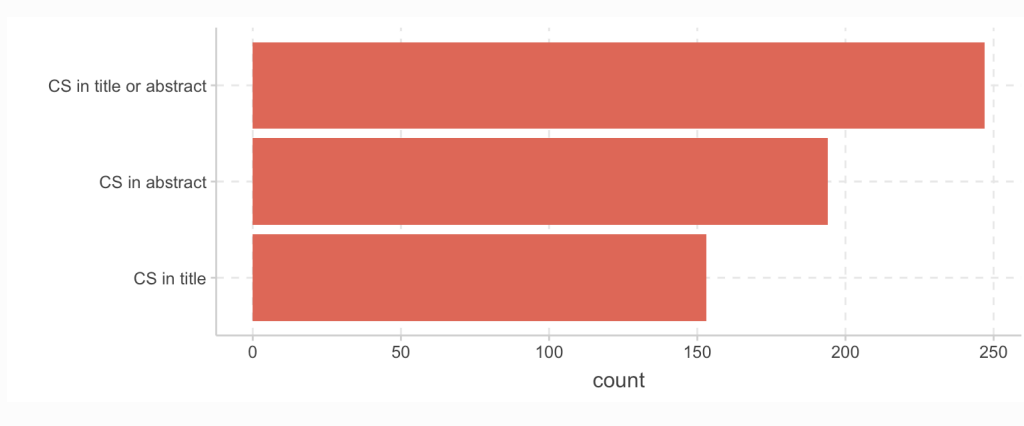
▼ Code

```
#create a new logical column with TRUE when AU-relevant terms mentioned in title s
dat$CS_title <- str_detect(string = str_to_lower(dat$Title), pattern =
"citizen|volunteer|crowdsourc")
#dat$Title[dat$CS_title == FALSE] lists that that did not match
# table(dat$CS_title)

#create a new logical column with TRUE when CS-relevant terms mentioned in abstracts
dat$CS_abs <- str_detect(string = str_to_lower(dat$Abstract.Note), pattern =
"citizen|volunteer|crowdsourc")
# table(dat$CS_abs)

#create a new logical column with TRUE when CS-relevant terms mentioned in titles or
abstracts
dat$CS_title_abs <- ifelse(dat$CS_title == "TRUE" | dat$CS_abs == "TRUE", "TRUE", "FALSE")
# table(dat$CS_title_abs)
```

Plot where citizen science(CS)-related terms are mentioned.

▼ Code

```
#summarise and merge for plotting for 3 new columns across all records
count_CS <- as.data.frame(rbind("CS in title" = table(dat$CS_title), "CS in abstract" =
table(dat$CS_abs), "CS in title or abstract" = table(dat$CS_title_abs)))
colnames(count_CS) <- c("not metioned", "mentioned")
count_CS$CS <- rownames(count_CS)

#plot
ggplot(count_CS, aes(x = reorder(CS, mentioned), y = mentioned)) + geom_bar(stat =
"identity") +
  coord_flip() +
  xlab("")  +
  ylab("count")
```



Authors

Process author infomation from Zotero dataset.

▼ Code

```
## Author column contains a mixture of different formats - try to Abbreviate author first
names

#split column with author names to find most common names
authors_all <- sort(unlist(str_split(dat$Author, "; "))) #mix of full and initialised first
names
#remove first word from each string (family name)
authors_all <- gsub(".*, ","", authors_all)
#extract next first word from each string (first name or initial)
authors_all <- stringr::word(authors_all, 1)
#check if string contains "." (use escape symbols!) and remove these strings
authors_all <- authors_all[!stringr::str_detect(authors_all, "\\.")]
count_authors_all <- as.data.frame(table(authors_all))

#remove numeric, comma, dot, single char, CSIRO and ID - in a new column
count_authors_all2 <- count_authors_all %>%
  mutate(authors_all2 = str_replace_all(authors_all, ".[:digit:].", "")) %>% #remove all
numbers
  mutate(authors_all2 = str_replace_all(authors_all2, ",", "")) %>%
  mutate(authors_all2 = str_replace_all(authors_all2, "ANTONY", "Antony")) %>%
  mutate(authors_all2 = str_replace_all(authors_all2, ">", "")) %>%
  subset(nchar(as.character(authors_all2)) > 1) %>% #remove 1-character strings
  subset(authors_all2 != "")  %>%  #remove empty strings
  mutate(name_length = str_length(authors_all2)) %>% #add column with name lengths
  arrange(desc(name_length)) %>% #arrange from the longest names to the shortest
  mutate(authors_abbr = str_replace_all(authors_all2, "[:lower:]", "")) %>% #remove all
lowercase characters and save in a new column (abbreviated names)
    mutate(authors_abbr = str_replace_all(authors_abbr, "-", "")) #remove hyphens

count_authors_all2 <- count_authors_all2[!count_authors_all2$authors_all %in%
c("Foundation", "Australian", "Electrical", "Foundation", "Department", "Terrestrial",
"Tecnologico", "Environmental", "GrrlScientist", "Contributing", "EcoCommons", "University",
"Computing", "Research", "Defence", "EPNRM", "SANBI", "ANPC", "EPPO", "OGTR", "TERN", "GSN",
"NSW", "OEH", "UNSW", "CSIRO", "this", "Water", "Olkola", "et", "an", "the", "The", "Joint",
"address=50", "Queensland", "Biodiversity"), ] #remove names that are not personal names

#authors_all_abbr <- setNames(object = count_authors_all2$authors_abbr,
count_authors_all2$authors_all) #create a named vector with patterns to be replaced to
abbreviate names
authors_all_abbr <- setNames(object = paste(",", count_authors_all2$authors_abbr),
paste(",", count_authors_all2$authors_all)) #create a named vector with patterns to be
replaced to abbreviate names, but obly for words following "," if more than one full first
name, only first one will be abbrevaietd (otherwise the family names get soetimes messed up)

dat$Author_abbr <- str_replace_all(dat$Author, authors_all_abbr) #replace authors first
names in the main data frame (as new column) with initials using the table above

#tidy up column with abbreviated first names (exclude names of groups and orgabisations)
dat$Author_abbr <- gsub("\\.", "", dat$Author_abbr) #remove dots
dat$Author_abbr <- gsub(" ", "", dat$Author_abbr) #remove white spaces
dat$Author_abbr <- gsub("scheme=AGLSTERMSAglsAgent", "", dat$Author_abbr)
dat$Author_abbr <- gsub("WatershedLandcare", "", dat$Author_abbr)
dat$Author_abbr <- gsub("AustralianGovernment-DepartmentofDefence,DefenceScience", "",
dat$Author_abbr)
dat$Author_abbr <- gsub("Technology,G", "", dat$Author_abbr)
dat$Author_abbr <- gsub("CanonInformationSystemsResearchAustraliaPty,L", "",
dat$Author_abbr)
dat$Author_abbr <- gsub("Griffith,University", "", dat$Author_abbr)
dat$Author_abbr <- gsub(";ElectronicsEngineersSignalProcessing,S", "", dat$Author_abbr)
dat$Author_abbr <-
gsub("Aprs;AustralianNational,University;IeeeComputerSociety;Nicta;TheComputerVision,Foundatio
"", dat$Author_abbr)
dat$Author_abbr <- gsub("CountryManagers,Y", "", dat$Author_abbr)
dat$Author_abbr <- gsub("Corporation,WAboriginal", "", dat$Author_abbr)
dat$Author_abbr <- gsub("NgukurrYangbalarangers", "", dat$Author_abbr)
dat$Author_abbr <- gsub("DirectorofNationalParks", "", dat$Author_abbr)
dat$Author_abbr <- gsub("TheInstituteof,Electrical", "", dat$Author_abbr)
dat$Author_abbr <- gsub("NativeConservationCouncil", "", dat$Author_abbr)
dat$Author_abbr <- gsub("DepartmentofEnvironment", "", dat$Author_abbr)
dat$Author_abbr <- gsub("TerrestrialEcosystems", "", dat$Author_abbr)
dat$Author_abbr <- gsub("CharlesSturtUniversity", "", dat$Author_abbr)
dat$Author_abbr <- gsub("NativeVegetationBranch", "", dat$Author_abbr)
dat$Author_abbr <- gsub("NSW,NConservationCouncil", "", dat$Author_abbr)
dat$Author_abbr <- gsub("address=50MarcusClarkeSt", "", dat$Author_abbr)
dat$Author_abbr <- gsub("EnvironmentalProtectionAuthority", "", dat$Author_abbr)
dat$Author_abbr <- gsub("OrganisationforEconomicCooperationandDevelopment", "",
dat$Author_abbr)
dat$Author_abbr <- gsub("NSWThreatenedSpeciesScientificCommittee", "", dat$Author_abbr)
dat$Author_abbr <- gsub("TheGreatEasternRangesInitiative(GER)", "", dat$Author_abbr)
dat$Author_abbr <-
gsub("AustralianGovernmentDeptofHealth:OfficeoftheGeneTechnologyRegulator", "",
dat$Author_abbr)
dat$Author_abbr <- gsub("DepartmentofAgriculture,WaterandtheEnvironment", "",
dat$Author_abbr)
dat$Author_abbr <- gsub("RylstoneRegionCoalFreeCommunity", "", dat$Author_abbr)
dat$Author_abbr <- gsub("SouthernHighlandNews", "", dat$Author_abbr)
dat$Author_abbr <- gsub("TERNEco-Informatics", "", dat$Author_abbr)
dat$Author_abbr <- gsub("EPNRMBoard", "", dat$Author_abbr)
dat$Author_abbr <- gsub("GrrlScientist", "", dat$Author_abbr)
dat$Author_abbr <- gsub("CanadensysNetwork", "", dat$Author_abbr)
dat$Author_abbr <- gsub("WarrnamboolCouncil", "", dat$Author_abbr)
dat$Author_abbr <- gsub("Tribune,TheNational", "", dat$Author_abbr)
dat$Author_abbr <- gsub("CityofJoondalup", "", dat$Author_abbr)
dat$Author_abbr <- gsub("NSWRoadsandMaritimeServices", "", dat$Author_abbr)
dat$Author_abbr <- gsub("RoyalBotanicGardensBoardVictoria", "", dat$Author_abbr)
dat$Author_abbr <- gsub("EnvironmentalProtectionAuthority", "", dat$Author_abbr)
dat$Author_abbr <- gsub("LitoriaEcoservices", "", dat$Author_abbr)
dat$Author_abbr <- gsub("DepartmentoftheEnvironment", "", dat$Author_abbr)
dat$Author_abbr <- gsub("DepartmentofAgricultureandWaterResources", "", dat$Author_abbr)
dat$Author_abbr <- gsub("NorthQueenslandBulkPortsCorporation", "", dat$Author_abbr)
dat$Author_abbr <- gsub("PhoenixEnvironmentalSciences", "", dat$Author_abbr)
dat$Author_abbr <- gsub("CSIRO", "", dat$Author_abbr)
dat$Author_abbr <- gsub("WADPW", "", dat$Author_abbr)
dat$Author_abbr <- gsub("SANBI", "", dat$Author_abbr)
dat$Author_abbr <- gsub("UNSW", "", dat$Author_abbr)
dat$Author_abbr <- gsub("Queensl,DRossDwyerRossDwyerisapost-
doctoralresearchfellowatTheUniversityof", "Dwyer,R", dat$Author_abbr)
dat$Author_abbr <- gsub("fundamental,Hisaresearcheracross", "", dat$Author_abbr)
dat$Author_abbr <- gsub("OzTrack,beenakeypartoftheteamdeveloping", "", dat$Author_abbr)
dat$Author_abbr <- gsub("ConselhoNacionaldeDesenvolvimentoCientificoe,Tecnologico", "",
dat$Author_abbr)
dat$Author_abbr <- gsub("SADepartmentforEnvironment,Water&NaturalResources", "",
dat$Author_abbr)
dat$Author_abbr <- gsub("NSWCatchmentManagementAuthority-NorthernRivers", "",
dat$Author_abbr)
dat$Author_abbr <- gsub("EuropeanCommission,JintResearchCentre", "", dat$Author_abbr)
dat$Author_abbr <-
gsub("AustralianNationalWildlifeCollection,NonalResearchCollectionsAustralia,,CanberraACT,A",
"", dat$Author_abbr)
dat$Author_abbr <- gsub("AustralianWildlifeConservancy,CairnsQLD4870,A", "",
dat$Author_abbr)
```

```
dat$Author_abbr <-
gsub("ResearchandRecoveryofEndangeredSpecies,SchoolofEarthandEnvironmentalSciences,Universityo
"", dat$Author_abbr)
dat$Author_abbr <- gsub("OlkolaAboriginalCorporation(Cairns)", "", dat$Author_abbr)
dat$Author_abbr <- gsub("corporateName=DepartmentofEducation,SandEmployment", "",
dat$Author_abbr)
dat$Author_abbr <- gsub("TheGreatEasternRangesInitiative(GER)", "", dat$Author_abbr)
dat$Author_abbr <- gsub("Linktoexternalsite,thislinkwillopeninanewwindow", "",
dat$Author_abbr)
dat$Author_abbr <- gsub("FisheriesQueensland,DepartmentofAgricultureandFisheries", "",
dat$Author_abbr)
dat$Author_abbr <-
gsub("CentreforEnvironmentalRiskManagementofBushfires,SchoolofEarth,AtmosphericandLifeSciences
"", dat$Author_abbr)
dat$Author_abbr <-
gsub("CentreforEcosystemScience,SchoolofBiologicalEarthandEnvironmentalSciences,UniversityofNe
"", dat$Author_abbr)
dat$Author_abbr <-
gsub("ResearchandRecoveryofEndangeredSpecies,SchoolofEarthandEnvironmentalSciences,Universityo
"", dat$Author_abbr)
dat$Author_abbr <- gsub("Ramirez-Cabral,NYakovlevaZitz", "Ramirez-Cabral,NYZ",
dat$Author_abbr)
dat$Author_abbr <- gsub("Platell,GAnneMarieMarguerite", "Platell,GAMM", dat$Author_abbr)
dat$Author_abbr <- gsub("daSilvaSantos,KCristineBezerra", "daSilvaSantos,KCB",
dat$Author_abbr)
dat$Author_abbr <- gsub("Ladino,PMiguelOliveira", "Ladino,PMO", dat$Author_abbr)
dat$Author_abbr <- gsub("Blanco,CMariaRodrigues", "Blanco,CMR", dat$Author_abbr)
dat$Author_abbr <- gsub("Mora-Cross,MAuxiliadora", "Mora-Cross,MA", dat$Author_abbr)
dat$Author_abbr <- gsub("SalazarCarrillo,JCarlos", "SalazarCarrillo,JC", dat$Author_abbr)
dat$Author_abbr <- gsub("Ezquivelzeta,ERobredo", "Ezquivelzeta,ER", dat$Author_abbr)
dat$Author_abbr <- gsub("Bickerstaff,JRonaldMunro", "Bickerstaff,JRM", dat$Author_abbr)
dat$Author_abbr <- gsub("Arias-Penna,DCarolina", "Arias-Penna,DC", dat$Author_abbr)
dat$Author_abbr <- gsub("Gerhardinger,LCavaleri" , "Gerhardinger,LC", dat$Author_abbr)
dat$Author_abbr <- gsub("AlmeydaZambrano,AMaria" , "lmeydaZambrano,AM", dat$Author_abbr)
dat$Author_abbr <- gsub("Wijewardhana,UAbhisheka" , "Wijewardhana,UA", dat$Author_abbr)
dat$Author_abbr <- gsub("Vasconcelos,HCristina" , "Vasconcelos,HC", dat$Author_abbr)
dat$Author_abbr <- gsub("Notarnicola,RFabrizio" , "Notarnicola,RF", dat$Author_abbr)
dat$Author_abbr <- gsub("([[:]])\\1+", "\\1", dat$Author_abbr)

#readr::write_csv(dat, file = here("data","dat_all_processed.csv")) #saving intermediate
version of the processed data
#dat <- read.csv(here("data", "dat_all_processed.csv")) #loading all data
```

Co-authorships

Count authors.

▼ Code

```
#all
#length(unlist(str_split(string = dat$Author_abbr, pattern = ";"))) # total number of
authors (might be slightly biased)
#length(unique(unlist(str_split(string = dat$Author_abbr, pattern = ";")))) # total number
of unique authors (might be slightly biased)
#table(dat$Author_abbr=="", dat$article_grey) #missing authorship (either not captured in
Zotero or removed organisations names)

#all - split and clean before counting again
authors_all <- unlist(str_split(string = dat$Author_abbr, pattern = ";"))
#hist(nchar(authors_all)) #check distribution of name lengths - some odd with >25 characters
#authors_all[nchar(authors_all) > 25]
authors_all <- authors_all[nchar(authors_all) < 26 & !is.na(authors_all)] #remove any odd or
empty long strings left
# hist(nchar(authors_all)) #check distribution of name lengths
# authors_all[nchar(authors_all) > 20]
# length(authors_all) # total number of author names (should be correct)
# length(unique(authors_all)) #  total number of unique authors names

#articles
authors_articles <- unlist(str_split(string = dat[dat$article_grey == "article",
"Author_abbr"], pattern = ";"))
authors_articles <- authors_articles[nchar(authors_articles) < 26 &
!is.na(authors_articles)]
# length(authors_articles) # total number of authors
# length(unique(authors_articles)) # total number of unique authors

#grey
authors_grey <- unlist(str_split(string = dat[dat$article_grey == "grey", "Author_abbr"],
pattern = ";"))
authors_grey <- authors_grey[nchar(authors_grey) < 26 & !is.na(authors_grey)]
# length(authors_grey) # total number of authors
# length(unique(authors_grey)) # total number of unique authors
```

Number of records with missing authorship information: 149 total, 3 for articles, 146 for grey literature.
Total number of unikwe known author names: 9813 total, 7827 for articles, 2742 for grey literature.

Author overlaps bytween literature types and more processing code.

▼ Code

```
## Authorship overlaps - are there same author names in articles and grey subsets?
# length(intersect(unique(authors_articles), unique(authors_grey))) /
length(unique(authors_all)) *100 # % authors in both articles and grey literature
# length(setdiff(unique(authors_grey), unique(authors_articles))) #count in grey but not in
articles - calculate precentages of authors publishing both
# length(setdiff(unique(authors_articles), unique(authors_grey))) #7071 in articles but not
in grey literature


# ## Multiple papers by the same author -  Top10 most productive authors for articles
# data.frame(table(authors_articles)) %>% rename(Names = authors_articles, Count = Freq) %>%
arrange(-Count) %>% slice_head(n = 10) %>%
# ggplot(aes(x = reorder(Names, Count), y = Count)) + geom_bar(stat = "identity") +
#   coord_flip() +
#   xlab("")  +
#   ylab("article count")
#

# ## Top10 most productive authors for grey
# data.frame(table(authors_grey)) %>% rename(Names = authors_grey, Count = Freq) %>%
arrange(-Count) %>% filter(Names != "") %>% slice_head(n = 10) %>%
# ggplot(aes(x = reorder(Names, Count), y = Count)) + geom_bar(stat = "identity") +
#   coord_flip() +
#   xlab("")  +
#   ylab("article count")

## Separate author names for overall summary per paper (list of vectors format)
Author_split <- str_split(dat$Author_abbr, ";") #list of vectors of authors
#length(Author_split) #3774 elements (outputs) in the list
dat$Author_counts <- lengths(Author_split) #numbers of elements in the vectors
#View(data.frame(dat$article_grey, dat$Author, dat$Author_abbr, dat$Author_counts)) #visual
checks
#table(dat$Author_counts, useNA = "always")
#hist(dat$Author_counts, breaks = 50) #most have just 1 author

#check for empty values
#table(dat$article_grey, dat$Author_abbr == "") #146/1356\8 (10%) of grey has missing
(individual) authorship, <1% for articles
#table(dat$article_grey, dat$Author_counts == "") #no empty values, i.e. missing ones count
as 1
dat$Author_counts <- replace(dat$Author_counts, dat$Author_abbr %in% c("", NA), 0) #replace
empty ones with 0
#table(dat$article_grey, dat$Author_counts == "0")
#dat[dat$Author_counts == 0, "Author_abbr"]

# ## Summarise author number per output by literature type
# dat %>%
#   group_by(article_grey) %>%
#   summarise(mean_AC = mean(Author_counts),
#           median_AC = median(Author_counts),
#           sd_AC = sd(Author_counts),
#           IQR_AC = IQR(Author_counts),
#           min_AC = min(Author_counts),
#           max_AC = max(Author_counts)
#           ) #does not capture that much of the grey lit is by teams, but under
organisational name

## Papers with multiple authors
# #create a dataframe of author counts and literature type
#  dat %>%
#    group_by(article_grey) %>%
#    reframe(article_grey, Author_counts) %>%
#    drop_na() %>% filter(Author_counts > 50) -> Author_counts_df
# table(Author_counts_df$article_grey) #number of papers with >50 authors

# dat %>%
#   reframe(article_grey, Title, Author_counts) %>%
#   filter(Author_counts > 50) #display paper titles with >50 authors
```

Authors that contributed to both articles and grey literature as % of the total number of unique author names: 7.7

Models for historical author team size changes for articles and grey literature.

▼ Code

```
# ## Author numbers per paper across years
# #create a data frame of author counts and literature type
dat %>%
  group_by(article_grey) %>%
    reframe(Publication.Year, Author_counts) %>%
  drop_na() -> count_year_authors

# #plot for articles
# count_year_authors %>%
#   filter(article_grey == "article") %>% # articles only
#   ggplot(aes(x = Publication.Year, y = Author_counts)) +
#   geom_point(alpha = 0.2, color = "#00cc99") +
#   ylim(0, 220) +
#   theme_minimal() +
#   theme(legend.position = "none") +
#   ylab("author count") +
#   geom_smooth(method = lm)
#
# #plot for grey
# count_year_authors %>%
#   filter(article_grey == "grey") %>% # grey only
#   ggplot(aes(x = Publication.Year, y = Author_counts)) +
#   geom_point(alpha = 0.2, color = "#000000") +
#   ylim(0, 220) +
#   theme_minimal() +
#   theme(legend.position = "none") +
#   ylab("author count") +
#   geom_smooth(method = lm)
#
# ggsave(plot = p, here("plots", "plot.authors_peroutput_years.pdf"), width = 16, height =
6, units = "cm",  dpi = "retina", scale = 2)

# #linear regression for articles with Year
model_a <- lm(Author_counts ~ scale(Publication.Year), data =
count_year_authors[count_year_authors$article_grey == "article" &
count_year_authors$Author_counts != 0, ]) #run for articles, without 0 values (missing data)
summary(model_a) #slope 0.7204      0.1484   4.854 1.29e-06 ***
```

```
Call:
lm(formula = Author_counts ~ scale(Publication.Year), data = count_year_authors[count_year_auth
    "article" & count_year_authors$Author_counts != 0, ])

Residuals:
   Min     1Q Median     3Q    Max
-4.902 -2.698 -1.217  0.579 212.339

Coefficients:
                        Estimate Std. Error t value Pr(>|t|)
(Intercept)               4.9051     0.1484  33.053  < 2e-16 ***
scale(Publication.Year)   0.7204     0.1484   4.854 1.29e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 7.275 on 2401 degrees of freedom
Multiple R-squared:  0.009716,  Adjusted R-squared:  0.009303
F-statistic: 23.56 on 1 and 2401 DF,  p-value: 1.291e-06
```

▼ Code

```
#
# #linear regression for grey with Year
model_g <- lm(Author_counts ~ scale(Publication.Year), data =
count_year_authors[count_year_authors$article_grey == "grey" &
count_year_authors$Author_counts != 0, ]) #run for grey, without 0 values (missing data)
summary(model_g) #slope 0.1411     0.1901   0.742    0.458
```

```
Call:
lm(formula = Author_counts ~ scale(Publication.Year), data = count_year_authors[count_year_auth
    "grey" & count_year_authors$Author_counts != 0, ])

Residuals:
   Min     1Q Median     3Q    Max
-2.078 -1.861 -1.645  0.355 208.009

Coefficients:
                        Estimate Std. Error t value Pr(>|t|)
(Intercept)               2.8556     0.1900  15.028   <2e-16 ***
scale(Publication.Year)   0.1411     0.1901   0.742    0.458
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 6.615 on 1210 degrees of freedom
Multiple R-squared:  0.0004551, Adjusted R-squared:  -0.000371
F-statistic: 0.5509 on 1 and 1210 DF,  p-value: 0.4581
```

▼ Code

```
## Stats for comparing number of authors per article vs. grey
dat %>%
  filter(Author_counts != 0) %>% #dont use 0 values (missing data)
    reframe(article_grey, Author_counts) -> Author_counts_df
# poisson - less likely to fit to real biological data well because of overdispersion
fit <- glm(Author_counts ~ article_grey, family = "poisson", data = Author_counts_df)
summary(fit)
```

```
Call:
glm(formula = Author_counts ~ article_grey, family = "poisson",
    data = Author_counts_df)

Coefficients:
                  Estimate Std. Error z value Pr(>|z|)
(Intercept)       1.590279   0.009211  172.65   <2e-16 ***
article_greygrey -0.540994   0.019333  -27.98   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 14645  on 3614  degrees of freedom
Residual deviance: 13787  on 3613  degrees of freedom
AIC: 24489

Number of Fisher Scoring iterations: 5
```

▼ Code

```
# #An alternative to overdispersed poisson fit - quasipoisson -> same pattern
# fit <- glm(Author_counts ~ article_grey, family = "quasipoisson", data = Author_counts_df)
# #An alternative to overdispersed poisson fit -  negative binomial - more likely to fit to
real biological data well -> same pattern
# fit <- MASS::glm.nb(Author_counts ~ article_grey, data = Author_counts_df)

#readr::write_csv(dat, file = here("data","dat_all_processed.csv")) #saving intermediate
version of the processed data
#dat <- read.csv(here("data", "dat_all_processed.csv")) #loading all data
```

Co-authorship network

Note: not executing the code below due to long time required to plot networks with so many nodes and connections.

Author networks calculations.

▼ Code

```
## Separate author names for overall summary per paper (list of vectors format) for article
Author_split <- str_split(string = dat[ , "Author_abbr"], pattern = ";") #list of vectors of
authors
length(Author_split) # elements (outputs) in the list

df_auth_collab <- data.frame(ref.id = rep(dat[, "Key"],
                             lengths(Author_split)),
                             auth_names = unlist(Author_split))
#table(df_auth_collab$ref.id == "") #no empty
#table(df_auth_collab$auth_names == "") # empty ""

#remove empty nodes
#df_auth_collab <- df_auth_collab[!is.na(df_auth_collab$auth_names), ]
df_auth_collab <- df_auth_collab[df_auth_collab$auth_names != "", ]
#length(unique(df_auth_collab$ref.id))
#length(unique(df_auth_collab$auth_names))

df_auth_collab %>%
  inner_join(df_auth_collab, by = "auth_names", relationship = "many-to-many") %>%
  distinct() %>% #remove many-to-many duplicates
  filter(ref.id.x <= ref.id.y) %>% #use <= to include unconnected outcomes (self-connections
counted as 1)
  count(ref.id.x, ref.id.y) %>%
  graph_from_data_frame(directed = FALSE) -> all_outputs_graph #with output ID (Key) names
as vertices
#str(all_outputs_graph)
```

Plot co-authorship network for articles and grey literature.
Points are items (articles/grey) and edges/connections are shared authors between published items.

▼ Code

```
# #add color attribute showing if article or grey
V(all_outputs_graph)$type <- ifelse(V(all_outputs_graph)$name %in% dat[dat$article_grey ==
"article", "Key"], "article", "grey")
# #table(V(all_outputs_graph)$color)
V(all_outputs_graph)$color <- ifelse(V(all_outputs_graph)$name %in% dat[dat$article_grey ==
"article", "Key"], "#0FD19D", "#808080")
# #table(V(all_outputs_graph)$color)

# #plot - each vertex is an output and edges reflect shared authors
# all_outputs_graph_layout <- layout_with_graphopt(all_outputs_graph)
#plot(all_outputs_graph)  #layout_nicely(all_outputs_graph)
#layout.kamada.kawai(all_outputs_graph)  # choose an appropriate graph layout algorithm for
the graph automatically
# all_outputs_graph_plot <- plot(all_outputs_graph, layout = all_outputs_graph_layout,
vertex.label = NA, vertex.size = 1, vertex.frame.color = NA)
# #, vertex.color = "grey"
# legend("left", legend=c("grey", "article"), col =
levels(as.factor(V(all_outputs_graph)$color)) , bty = "n", pch = 20 , pt.cex = 3, cex = 1.5,
text.col="black" , horiz = FALSE, inset = c(0.1, 0.1))

type <- V(all_outputs_graph)$type #save as an object for easier plotting

## Plot graph using library(ggraph) - outcomes as nodes (single-author articles shown)
#https://www.r-bloggers.com/2020/03/ggraph-tricks-for-common-problems/

p_network <- ggraph(all_outputs_graph, 'igraph', algorithm = 'kk') +   #, colour =
'forestgreen'
    #geom_edge_link0(aes(width = 0.01), edge_alpha = 0.1) +
    geom_edge_fan(color = "gray50", width = 0.5, alpha = 0.2) +
    geom_node_point(aes(linewidth = 0.001, fill = type, color = type, alpha = 0.01)) + #
using directly $color for geom_node_point does not work
    scale_colour_manual(breaks = c("article", "grey"), values = c("#66c2a5", "#080808")) +
    guides(alpha = "none", size = "none", type = "none") + #remove unnecessary legend elements
    labs(colour = "type") + #change legend title
    theme_void()
```

Collaboration networks

Authors are nodes and edges/connections are items they published together.
Again, not executing the plots due to the large size of the networks.

Collaboration networks (using igraph) - articles subset.

▼ Code

```
### article literature subset
## Separate author names for overall summary per paper (list of vectors format) for article
Author_split_article <- str_split(string = dat[dat$article_grey == "article",
"Author_abbr"], pattern = ";") #list of vectors of authors
length(Author_split_article) #2406 elements (outputs) in the list

df_auth_collab_article <- data.frame(ref.id = rep(dat[dat$article_grey == "article", "Key"],
lengths(Author_split_article)),
                      auth_names = unlist(Author_split_article))
#str(df_auth_collab_article)
#df_auth_collab_article[1:20, ]

#remove empty ones
df_auth_collab_article <- df_auth_collab_article[!is.na(df_auth_collab_article$auth_names),
]
df_auth_collab_article <- df_auth_collab_article[df_auth_collab_article$auth_names != "", ]

#find 25 most productive authors
# df_auth_collab_article %>% count(auth_names) %>% arrange(-n) %>% slice_head(n = 25)

#compute graph of collaborations for all authors
df_auth_collab_article %>%
  inner_join(df_auth_collab_article, by = "ref.id", relationship = "many-to-many") %>%
  distinct() %>% #remove many-to-many duplicates
  filter(auth_names.x < auth_names.y) %>%
  count(auth_names.x, auth_names.y) %>%
  graph_from_data_frame(directed = FALSE) -> a1 #with author names as vertices

summary(a1) #IGRAPH 2b42183 UN-- 7650 166630 -- attr: name (v/c), n (e/n)
#an undirected multigraph (parallel edges) with 7650 authors and 166630 scientific
collaborations. Each node (author) in the network has 1 attribute: name. Each edge has 1
attributes: n (collaborations).

#overall graph statistics
graph.density(a1) #0.006 = quite sparse
transitivity(a1, type = "global") # 0.894 = probability that the neighbors of a node are
also connected (also called the clustering coefficient)

#see components of the igraph object
# E(a1)
# V(a1)
# a1[]
# edge_attr(a1)
# vertex_attr(a1)
# as_data_frame(a1, what = "edges")
# get.edgelist(a1)
```

Plot article author collaboration network.
Points are authors and edges/connections are shared articles they published together.
(Note: not plotting due to this taking long time - the graph is very dense!)

▼ Code

```
#visualize our co-authorship graph network
(https://stackoverflow.com/questions/9759484/vertex-border-color-width-in-r-graph-plot)
a1_graph_layout <- layout_with_graphopt(a1) #layout_nicely(a1) #layout.kamada.kawai(a1)  #
choose an appropriate graph layout algorithm for the graph automatically
a1_graph_plot <- plot(a1, layout = a1_graph_layout, vertex.label = NA, vertex.size = 1,
vertex.color = "#66c2a5")
```

Collaboration network graph metrics from a saved network object - articles.
(Can be run manually from a saved data object)

▼ Code

```
load(here("data", "data_a1.RData")) #load saved data object

#try to simplify the graph
a1is <- simplify(a1, remove.multiple = T, remove.loops = T,
                 edge.attr.comb=list(weight="sum", "ignore") ) #simplify

# plot(a1is, layout = a1_graph_layout, edge.arrow.size=0, vertex.color="gold",
vertex.size=1, vertex.frame.color="gray",
#       vertex.label.color="black", vertex.label.cex=0.8, vertex.label.dist=2,
edge.curved=0.2, vertex.label = NA)
# #E(a1is)
# #V(a1is)

##Find cliques (complete subgraphs of an undirected graph) - not good - picks up the 209-
authors paper
# cliques(as.undirected(a1)) # compute list of cliques
# sapply(cliques(as.undirected(a1)), length) # clique sizes
# largest_cliques(as.undirected(a1)) # clique with max number of nodes - 2010 - single paper

# vcol <- rep("article80", vcount(as.undirected(a1)))
# vcol[unlist(largest_cliques(as.undirected(a1)))] <- "gold"
# plot(as.undirected(a1), vertex.label=V(a1)$name, vertex.color=vcol)


#Calculating more centrality measures (degree, betweeness, closeness, eigenvector
centrality, coreness; see https://ona-book.org/vertex-importance.html and
https://users.dimi.uniud.it/~massimo.franceschet/R/communities.html)

metrics_article <- data.frame(type = rep("article", length(igraph::degree(a1is))),
  deg = igraph::degree(a1is),
  bet = igraph::betweenness(a1is),
  clo = igraph::closeness(a1is),
  eig = igraph::eigen_centrality(a1is)$vector,
  cor = igraph::coreness(a1is)
  )
cor.test(metrics_article$deg, metrics_article$clo) #clo signif correlated with deg
#summary(metrics_article$bet)
#hist(metrics_article$bet, breaks = 100) # mostly extremely low values and a few high ones


## Greedy community detection
# greedy method (hiearchical, fast method)
comm_a1 <- cluster_fast_greedy(a1is)
# modularity measure
modularity(comm_a1) #0.76
# memberships of nodes
#membership(comm_a1)
# number of communities
length(comm_a1)
# size of communities
#hist(sizes(comm_a1), breaks = 100) #mostly very small, but a few really large ones

# plot communities with shaded regions
#plot(a1c, a1, layout = a1_graph_layout, vertex.label = NA, vertex.size = 2)
# plot communities without shaded regions
#plot(a1, vertex.color = membership(c1), layout = a1_graph_layout, vertex.label = NA,
vertex.size = 2)

#save network for articles
save(a1, file = here("data", "data_a1.RData"))
# To load the data again
#load(here("data", "data_a1.RData"))
```

Collaboration networks (using igraph) - grey literature subset.

▼ Code

```
### Grey literature subset
## Separate author names for overall summary per paper (list of vectors format) for grey
Author_split_grey <- str_split(string = dat[dat$article_grey == "grey", "Author_abbr"],
pattern = ";") #list of vectors of authors
length(Author_split_grey) #1358 elements (outputs) in the list

df_auth_collab_grey <- data.frame(ref.id = rep(dat[dat$article_grey == "grey", "Key"],
lengths(Author_split_grey)),
                  auth_names = unlist(Author_split_grey))

#remove empty ones
df_auth_collab_grey <- df_auth_collab_grey[!is.na(df_auth_collab_grey$auth_names), ]
df_auth_collab_grey <- df_auth_collab_grey[df_auth_collab_grey$auth_names != "", ]

#find 25 most productive authors
# df_auth_collab_grey %>% count(auth_names) %>% arrange(-n) %>% slice_head(n = 25)

#compute graph of coauthorships for all authors
df_auth_collab_grey %>%
  inner_join(df_auth_collab_grey, by = "ref.id", relationship = "many-to-many") %>%
  distinct() %>% #remove many-to-many duplicates) %>%
  filter(auth_names.x < auth_names.y) %>%
  count(auth_names.x, auth_names.y) %>%
  graph_from_data_frame(directed = FALSE) -> g1 #with author names as vertices

summary(g1) #IGRAPH b8acb1a UN-- 2235 57722 -- attr: name (v/c), n (e/n)
#an undirected multigraph (parallel edges) with 2235 authors and 57722 scientific
collaborations. Each node (author) in the network has 1 attribute: name. Each edge has 1
attributes: n (collaborations).

#overall graph statistics
graph.density(g1) #0.023 = quite sparse
transitivity(g1, type = "global") # 0.993 = probability that the neighbors of a node are
also connected (also called the clustering coefficient)

#see components of the igraph object
# E(g1)
# V(g1)
# g1[]
# edge_attr(g1)
# vertex_attr(g1)
# as_data_frame(g1, what = "edges")
# get.edgelist(g1)

#save network for articles
save(g1, file = here("data", "data_g1.RData"))
# To load the data again
#load(here("data", "data_g1.RData"))
```

Plot grey author collaboration network.
Points are authors and edges/connections are shared articles they published together.
(Note: not plotting due to this taking long time - the graph is very dense!)

▼ Code

```
#load(here("data", "data_g1.RData")) #load saved data object
#visualize our co-authorship graph network
g1_graph_layout <- layout_with_graphopt(g1) #layout_nicely(g1) #layout.kamada.kawai(g1)  #
choose an appropriate graph layout algorithm for the graph automatically
g1_graph_plot <- plot(g1, layout = g1_graph_layout, vertex.label = NA, vertex.size = 1,
vertex.color = "grey")
```

Collaboration network graph metrics.
(Can be run from a saved network object.)

▼ Code

```
load(here("data", "data_g1.RData")) #load saved data object

#try to simplify the graph
g1is <- simplify(g1, remove.multiple = T, remove.loops = T,
                 edge.attr.comb=list(weight="sum", "ignore") ) #simplify

# plot(g1is, layout = g1_graph_layout, edge.arrow.size=0, vertex.color="gold",
vertex.size=1, vertex.frame.color="gray",
#       vertex.label.color="black", vertex.label.cex=0.8, vertex.label.dist=2,
edge.curved=0.2, vertex.label = NA)
# #E(g1is)
# #V(g1is)

##Find cliques (complete subgraphs of an undirected graph) - not good - picks up the 209-
authors paper
# cliques(as.undirected(g1)) # compute list of cliques
# sapply(cliques(as.undirected(g1)), length) # clique sizes
# largest_cliques(as.undirected(g1)) # clique with max number of nodes - 2010 - single paper

# vcol <- rep("grey80", vcount(as.undirected(g1)))
# vcol[unlist(largest_cliques(as.undirected(g1)))] <- "gold"
# plot(as.undirected(g1), vertex.label=V(g1)$name, vertex.color=vcol)


#Calculating more centrality measures (degree, betweeness, closeness, eigenvector
centrality, coreness)
metrics_grey <- data.frame(type = rep("grey", length(igraph::degree(g1is))),
  deg = igraph::degree(g1is),
  bet = igraph::betweenness(g1is),
  clo = igraph::closeness(g1is),
  eig = igraph::eigen_centrality(g1is)$vector,
  cor = igraph::coreness(g1is)
  )

cor.test(metrics_grey$deg, metrics_grey$clo) #clo signif correlated with deg
#summary(metrics_grey$bet)
#hist(metrics_grey$bet, breaks = 100) # mostly extremely low values and a few high ones

# ggplot(metrics_grey, aes(x = deg, fill = type)) +
#   geom_density(alpha = .3)

# ### STATS  - COMPARE a1 and g1 graph degree centrality metrics
# metrics_all <- rbind(metrics_article, metrics_grey) #create a single data frame
# str(metrics_all)
# fit_metrics_all <- glm(deg ~ type, family = "poisson", data = metrics_all) #node degree
centrality
# summary(fit_metrics_all) #grey slightly higher degree


## Greedy community detection
# greedy method (hiearchical, fast method)
comm_g1 <- cluster_fast_greedy(g1is)
# modularity measure
modularity(comm_g1) #0.399
# memberships of nodes
#membership(comm_g1)
# number of communities
length(comm_g1)
# size of communities
# sizes(comm_g1)

# COMPARE plot communities with shaded regions
# plot(comm_g1, g1is, layout = g1_graph_layout, vertex.label = NA, vertex.size = 2)
# plot communities without shaded regions
# plot(g1is, vertex.color = membership(comm_g1), layout = g1_graph_layout, vertex.label =
NA, vertex.size = 2)


## Extract top4 components
#str(g1)
#comps_grey <- components(g1)
#groups(comps_grey)

# #alt way to see components:
# comps_grey <- decompose(g1, min.vertices=2)
# sapply(comps_grey, gorder) #counts number of vertices of each component
# sapply(comps_grey, gsize) #counts number of edges of each group
# sapply(comps_grey, diameter) #calculates diameter of each group

# #a series of plots by component
# sapply(decompose(g1), plot)
# sapply(decompose(g1)[2], plot) #only second one
# sapply(decompose(g1)[2], vertex_attr)  #only second one, with names

# comps_grey <- components(g1)
# -sort(desc(comps_grey$csize)) #show descending sizes
# table(comps_grey$csize < 6) #70% of components have 5 or less authors
# -sort(desc(comps_grey$csize))[1:4] #sizes of the top 3 components
# comp_max <- which(c$membership == which.max(c$csize)) # get the largest component group
# comp_985 <- which(comps_grey$membership == which(comps_grey$csize == 985)) # get the
largest component group
# comp_31 <- which(comps_grey$membership == which(comps_grey$csize == 31)) # get the largest
component group
# comp_26 <- which(comps_grey$membership == which(comps_grey$csize == 26)) # get the largest
component group
# comp_23 <- which(comps_grey$membership == which(comps_grey$csize == 23)) # get the largest
component group

#compute centrality measures for vertexes - degree (number of links to a given author)
g1_degree <- degree(g1)
#hist(g1_degree) #a small number of high-degree vertexes - highly skewed (209 authors from
the same paper)
g1_bet <- betweenness(g1) #author betweenness measures
#hist(g1_bet) #highly skewed
authors_bet <- g1_bet[g1_bet > 25000] #22 with betweenes score >25000:
length(names(authors_bet))
#sort(authors_bet, decreasing = TRUE)
df_authors_bet <- df_auth_collab_grey[df_auth_collab_grey$auth_names %in%
names(authors_bet), ] #only keep ones that match names in authors_central_table, use ref.id
to see realted publications
#length(unique(df_authors_bet$ref.id)) #78 unique publications

df_authors_bet %>%
  inner_join(df_authors_bet, by = "ref.id", relationship = "many-to-many") %>%
  distinct() %>% #remove many-to-many duplicates
  filter(auth_names.x < auth_names.y) %>%
  count(auth_names.x, auth_names.y) %>%
  graph_from_data_frame(directed = FALSE) -> g1b #with author names as vertices

summary(g1b) #IGRAPH a3211e2 UN-- 23 42 --  + attr: name (v/c), n (e/n)
#overall graph statistics
graph.density(g1b) #0.166 = quite sparse
transitivity(g1b, type = "global") # 0.457 = prob
```

```
= probability that the neighbors of a node are
also connected (also called the clustering coefficient)

#see components of the igraph object
# E(g1b)
# V(g1b)
# g1b[]
# edge_attr(g1b)
# vertex_attr(g1b)
# as_data_frame(g1b, what = "edges")
# get.edgelist(g1b)

#visualize our co-authorship graph network
#g1b_graph_layout <- layout_with_graphopt(g1b) #layout_as_tree(g1b) #layout_on_sphere(g1b)
#layout_in_circle(g1b) #layout.kamada.kawai(g1b) #layout_nicely(g1b)
#plot(g1b, layout = g1b_graph_layout, vertex.size = betweenness(g1b), vertex.color = "grey",
vertex.frame.color="grey") #vertex.label = NA,

save(g1, file = here("data", "data_g1.RData")) # save network for articles
#load(here("data", "data_g1.RData")) # to load the data again
```

Summary of results from network analyses: We note that author team size potentially influences the 'connectedness' of the co-authorship network among the ALA-related outputs analysed above. The plotted network graph has nodes (vertices) representing outputs of two types (article, grey) which are connected via the author identity (edges). Formal graph analysis indicated that the co-authorship network of grey outputs had a relatively higher proportion of weakly connected items than an equivalent network of journal articles. A community detection algorithm revealed lower modularity for a grey outputs graph in comparison to journal articles (0.40 vs. 0.76) indicating that clustered journal articles are more tightly connected to each other by co-authorship, potentially via multiple shared authors.

Separate subsets for peer-reviewed (articles) and grey literature and seva intermittent files

Subset main data set.

▼ Code

```
dat_articles <- dat %>% filter(Item.Type == "journalArticle") %>%
dplyr::select(Key:Manual.Tags)
#dim(dat_articles) #2406
dat_grey <- dat %>% filter(Item.Type != "journalArticle") %>% dplyr::select(Key:Manual.Tags)
#dim(dat_grey) #1358

## Save main processed data and subsets:
readr::write_csv(dat, file = here("data","dat_all_processed.csv"))
readr::write_csv(dat_articles, file = here("data","dat_articles_processed.csv"))
readr::write_csv(dat_grey, file = here("data","dat_grey_processed.csv"))
#dat <- read.csv(here("data", "dat_all_processed.csv")) #loading all data
```

Author affiliations - articles

Zotero records do not contain information on authors affiliations.
Trying to fetch bibliographic records from OpenAlex API using article DOI.

OpenAlex

Get affiliation information for journal articles (via doi) from OpenAlex.org platform using openalexR package.
Note: do not run - skip this chunk and load locally saved fetched records.

▼ Code

```
dois <- unique(dat_articles$DOI) #some values need cleaning, as not proper DOI (empty
strings) or extra chars
dois <- gsub("https://doi.org/", "", dois)
dois <- dois[dois != ""]
dois <- dois[dois != "--"]
length(dois) #2071

#A custom function allowing to fetch multiple records (usually works for up to 500 before
throwing errors randomly) and convert the to bibliometrix-like format using oa2bibliometrix
block <- function(i){
  print(paste("fetch request nr ", i))
  Sys.sleep(0.2)
  fetched <- oa_fetch(
  doi = dois[i],
  entity = "works",
  verbose = TRUE)
  print(length(fetched))
#  return(fetched)
  if (length(fetched) > 0) oa2bibliometrix(fetched) else print("- not found") #convert or
print message
}

#use the custom functions to run fetch with pauses and save individual results as a list of
data frames - in chunks of around 500 records
block_test1 <- lapply(1:length(dois[1:500]), block)
block_test2 <- lapply(1:length(dois[501:1000]), block)
block_test3 <- lapply(1:length(dois[1001:1500]), block)
block_test4 <- lapply(1:length(dois[1501:2071]), block)
#note that json errors are thrown randomly and longer chunks are more likely to halt with #
Error: lexical error: invalid char in json text

#join block_test 1 to 4
block_test <- c(block_test1, block_test2, block_test3, block_test4)

#class(block_test) #list
length(block_test) #2071
table(lengths(block_test)) #2063 records with 44 elements, 8 records with 1 element (empty)

save(block_test, file = here("data", "results_OpenAlex_full.RData")) #save OpenAlex output
as a Rdata object
```

Processing affiliation country information. (Can be run from a saved data object).

▼ Code

```
load(file = here("data", "results_OpenAlex_full.RData")) #loads block_test data object with
full OpenAlex block query output
#extract two variables from the results
res_OpenA <- lapply(block_test, "[", c("doi", "AU_CO"))
  #data.frame(doi = sapply(block_test, "[", c("doi")), AU_CO = sapply(block_test, "[",
c("AU_CO")) )
#str(res_OpenA)
res_OpenA <- as.data.frame(do.call(rbind, res_OpenA))
#str(res_OpenA) #'data.frame':  2071 obs. of  2 variables: doi and AU_CO

## Processing affiliation country information into a data frame
doi_country_df <- data.frame(doi = tolower(gsub("https://doi.org/", "", res_OpenA$doi)),
AU_CO = res_OpenA$AU_CO)
# table(is.na(doi_country_df)) #16/2 = 8 empty records
# doi_country_df[is.na(doi_country_df), ]
doi_country_df <- doi_country_df[!is.na(doi_country_df$doi), ] #remove empty records
#dim(doi_country_df) #2063 records = articles
doi_country_list <- strsplit(doi_country_df$AU_CO, ";") #creates a list of vectors of
country names - many are NA
#length(doi_country_list) #2063
doi_country_list <- lapply(doi_country_list, setdiff, 'NA') #removes country names that are
NA
doi_country_list <- doi_country_list[lengths(doi_country_list) > 0] #remove list elements
that are empty - character(0)
#length(doi_country_list) #1765 with non-NA country names, i.e. have any usable country info
#sum(lengths(doi_country_list)) #2681 number of authors with affiliation country (after
excluding NA values)
#table(lengths(lapply(doi_country_list, unique)))  #table of counts of  unique country names
per article (after excluding NA values)
#1235/1765 #70% of articles with single-country authorship

#How many papers with only Australian authors?
##################View(doi_country_list[lengths(lapply(doi_country_list, unique)) == 1])
doi_country_list2 <- sapply(doi_country_list, unique) #unique country names
table(unlist(doi_country_list2[lapply(doi_country_list2, length) == 1]) == "AUSTRALIA") #
Filter out elements with length 0, unlist and then check which are "AUSTRALIA" = 1030/1765
#58% of articles with only Australian authorship
```

```
FALSE  TRUE
  205  1030
```

▼ Code

```
doi_country_all <- unlist(strsplit(doi_country_df$AU_CO, ";")) #extract country names as a
single vector
#length(doi_country_all) #, including NA
#table(doi_country_all) #table of country frequencies (note 2098 NA = missing)
#dim(table(doi_country_all)) #54 different country names (including NA)
doi_country_all <- doi_country_all[doi_country_all != "NA"] #remove NA
#length(doi_country_all) #total 8672 author country values after removing NA
```

There are 1765 articles with author country information (for 2681 authors).
Overall, authors are affiliated with 53 countries.
List of top 10 countries in terms of the numbers of affiliated authors:

▼ Code

```
##Count number of authors per country
country_count <- data.frame(table(doi_country_all))
#str(country_count)
country_count <- arrange(country_count, -Freq)
country_count$percent <- round(country_count$Freq/sum(country_count$Freq)*100, 1)
top10countries <- country_count[1:10,]
top10countries
```

```
   doi_country_all Freq percent
1        AUSTRALIA 5880    67.8
2              USA  695     8.0
3          GERMANY  329     3.8
4   UNITED KINGDOM  200     2.3
5     SOUTH AFRICA  166     1.9
6           BRAZIL  138     1.6
7            SPAIN  130     1.5
8      NEW ZEALAND  127     1.5
9           CANADA   91     1.0
10          SWEDEN   77     0.9
```

Country-level collaboration network

Process country of affiliation data - create collaboration matrices.
(Output not shown due to very long processing time).

▼ Code

```
#create big matrix of coauthorship links among all countries (incl NA)
doi_country_df$DB <- "SCOPUS" #add extra column to enable processing with biblioNetwork
function below
NetMatrix <- biblioNetwork(doi_country_df, analysis = "collaboration", network =
"countries", sep = ";")
#str(NetMatrix)
net_matrix <- as.matrix(NetMatrix)
#str(net_matrix)
#country_sum <- data.frame(countries = rownames(net_matrix), sum = rowSums(net_matrix)) #sum
of links per country

#use only top 10 countries to create a smaller matrix
small_matrix <- net_matrix[rownames(net_matrix) %in% top10countries$doi_country_all,
colnames(net_matrix) %in% top10countries$doi_country_all] #reduce matrix to most productive
countries (with highest total number of authors)
diag(small_matrix) <- 0 #get rid of collaboration counts within same country

# getting rid of lower triangle (as this is duplication of info)
small_matrix[lower.tri(small_matrix)] <- 0
#colnames(small_matrix) - change to title case:
colnames(small_matrix) <- str_to_title(colnames(small_matrix))
#rownames(small_matrix) - change to title case:
rownames(small_matrix) <- str_to_title(rownames(small_matrix))
#Fix abbreviated country names :
colnames(small_matrix)[colnames(small_matrix) == "Usa"] <- "USA"
rownames(small_matrix)[rownames(small_matrix) == "Usa"] <- "USA"
colnames(small_matrix)[colnames(small_matrix) == "United Kingdom"] <- "United Kingdom"
rownames(small_matrix)[rownames(small_matrix) == "United Kingdom"] <- "United Kingdom"
```

Article country collaboration chord network plot.

▼ Code

```
##Plot country collaborations as a chord diagram for top 10 countries (using circlize
package).
##Circumference are countries and edges/connections are shared articles they published
together.
circos.clear()
circos.par(canvas.xlim = c(-1, 1), canvas.ylim = c(-1.2, 1.5)) #set margins
#pdf(file = here("plots", "Figure_country_collaboration.pdf"), width=8, height=8,
pointsize=10)
chordDiagramFromMatrix(small_matrix) #actual plotting
#dev.off()
```

Plot chord diagram to pdf.

▼ Code

```
pdf(file = here("plots", "Fig4_rev_plot.chord_country_coauthorship.pdf"), width=8, height=8,
pointsize=12)
par(mar=c(0,0,0,0))
par(oma=c(0,0,0,0))
circos.clear()
circos.par(canvas.xlim = c(-1, 1), canvas.ylim = c(-1.2, 1.5)) #set margins
chordDiagramFromMatrix(small_matrix) #actual plotting
dev.off()
```

Plot on the world map and save to pdf.

▼ Code

```
##load and filter world map data
world_map <- map_data("world") %>%
  filter(! long > 180) #remove countries with longitude >180 to make equal projection-like
map without artifacts
#table(world_map$region) #note that United Kingdom is UK and America is USA
#table(world_map$region) #note that United Kingdom is UK and America is USA

##format country names to match regions on the world map
country_count$region <- str_to_title(as.character(country_count$doi_country_all)) #create
title case names for matching
country_count$region[country_count$region == "Usa"] <- "USA" #fix
country_count$region[country_count$region == "United Kingdom"] <- "UK" #fix
country_count$region[country_count$region == "United Kindom"] <- "UK" #fix
country_count$region[country_count$region == "Russian Federation"] <- "Russia"
country_count$region[country_count$region == "Czechia"] <- "Czech Republic"
country_count$region[country_count$region == "Bolivia (Plurinational State Of)"] <-
"Bolivia"
country_count$region[country_count$region == "Venezuela (Bolivarian Republic Of)"] <-
"Venezuela"
#country_count$region[country_count$region == "Hong Kong"] <- "China" #this messess up the
counts?
country_count$region[country_count$region == "Viet Nam"] <- "Vietnam"
country_count$region[country_count$region == "Brunei Darussalam"] <- "Brunei"
country_count$region[country_count$region == "Congo"] <- "Republic of Congo"
country_count$region[country_count$region == "Syrian Arab Republic"] <- "Syria"
country_count$region[country_count$region == "Tanzania, United Republic Of"] <- "Tanzania"

#country_count$region %in% world_map$region #check matching
#(country_count$region)[!(country_count$region %in% world_map$region)] #check matching

## colour all regions on the map:
emptymap <- tibble(region = unique(world_map$region), n =
rep(0,length(unique(world_map$region)))) #create table with all counts as 0
fullmap <- left_join(emptymap, country_count, by = "region") #join with actual counts table
fullmap$count <- fullmap$n + fullmap$Freq # make new column for fixed counts
fullmap$count[is.na(fullmap$count)] <- 0 #change NA to 0 for regions with no counts
##plot
fullmap %>%
  ggplot(aes(fill = count, map_id = region)) +
  geom_map(map = world_map) +
  expand_limits(x = world_map$long, y = world_map$lat) +
  coord_map("moll") +
  ggthemes::theme_map() +
  theme(legend.position="right") +
  scale_fill_gradient(low = "lightgrey", high = "darkgreen", name = "author count",
      guide = guide_colorbar(direction = "vertical.")) +
  guides(fill = guide_colourbar(barwidth = unit(5, units = "mm"), barheight = unit(30, units
= "mm")))

#save the figure to pdf
#ggsave(here("plots", "plot.map_author_counts.pdf"), width = 16, height = 6, units = "cm",
dpi = "retina", scale = 2)
```

Articles - journals names

Summarise journal-level information - journal names.

Clean journal names.

▼ Code

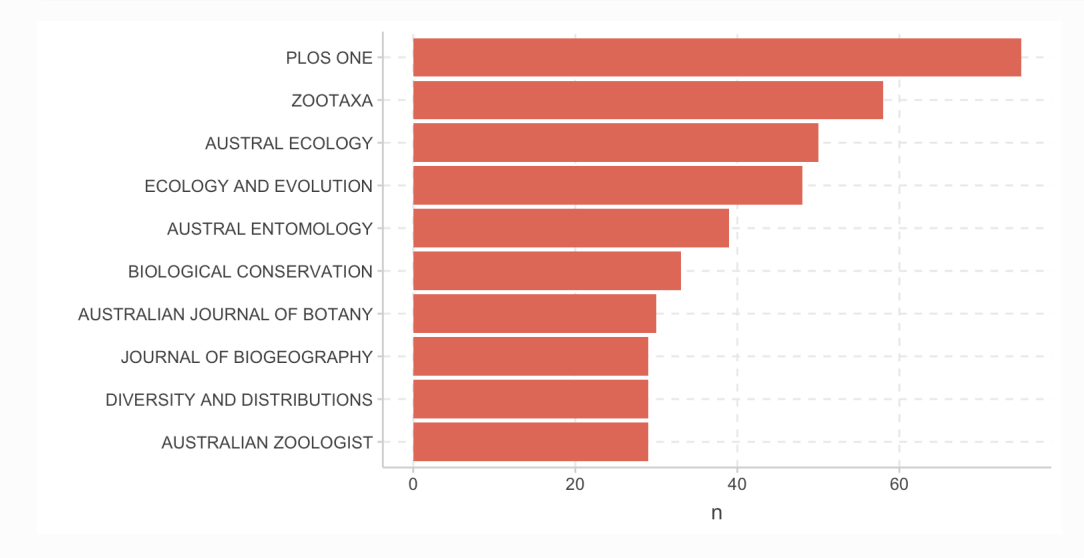
```
#length(unique((dat_articles$Publication.Title))) #774 - some names duplicated due to
different capitalisation
#length(unique(str_to_upper(dat_articles$Publication.Title))) #759 when all uppercase
dat_articles$Journal_name <- str_to_upper(dat_articles$Publication.Title) #create a new
column with uppercase journal names
# View(table(dat_articles$Journal_name)) #visual check
dat_articles$Journal_name <- gsub("&", "AND", dat_articles$Journal_name, fixed = TRUE)
#replace all & with AND
dat_articles$Journal_name <- gsub("EMU - AUSTRAL ORNITHOLOGY", "EMU",
dat_articles$Journal_name, fixed = TRUE) #replace EMU - AUSTRAL ORNITHOLOGY with EMU
#length(unique(dat_articles$Journal_name)) #753
#table(dat_articles$Journal_name == "") #15 missing Journal_name
#View(dat_articles[dat_articles$Journal_name == "", ]) #see records missing journal name -
some are from ResearchGate records, local journals or conference publications
#dat_articles[dat_articles$Journal_name == "", "Title"] #titles of papers missing journal
name
```

The articles were published in 753 different journals.

Plot top 10 most frequent publication journals.

▼ Code

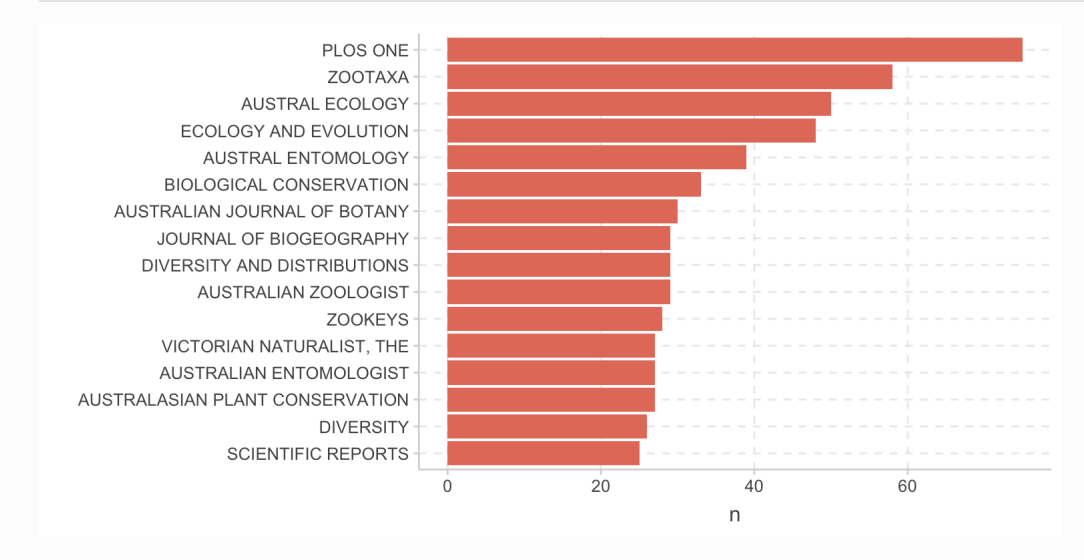
```
dat_articles %>%
  count(Journal_name) %>% slice_max(order_by = n, n = 10) ->  count_Journal_name #get top10
most common journal names

ggplot(count_Journal_name, aes(x = reorder(Journal_name, n), y = n)) +
  geom_bar(stat = "identity") +
  coord_flip() +
  xlab("")
```



Plot journals with at least 25 articles.

▼ Code

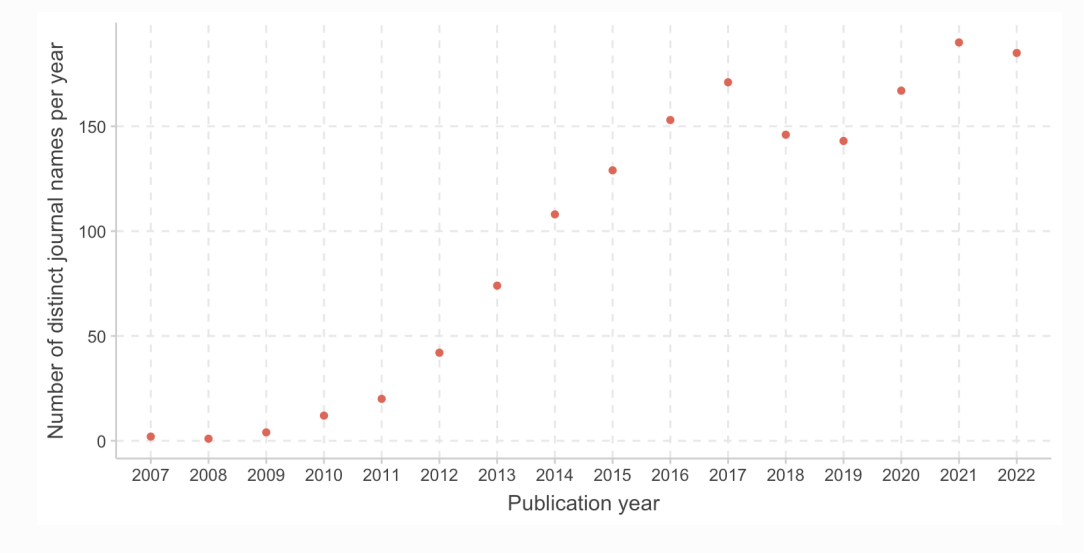
```
dat_articles %>%
  count(Journal_name) %>% filter(n >= 25) -> count_Journal_name #getjournal names with at
least 25 articles

ggplot(count_Journal_name, aes(x = reorder(Journal_name, n), y = n)) +
  geom_bar(stat = "identity") +
  coord_flip() +
  xlab("")
```



Plot number of distinct publication journals per publication year.

▼ Code

```
#plot number of distinct journal names per year
dat_articles %>%
  group_by(Publication.Year) %>%
  summarise(n = n_distinct(Journal_name)) %>% ggplot(
      aes(x = str_to_title(Publication.Year),
          y = n)) +
  geom_point() +
  xlab("Publication year") +
  ylab("Number of distinct journal names per year")
```



Plot trends of number of articles in most represented journals per publication year.

▼ Code

```
dat_articles %>%
  count(Journal_name, Publication.Year) %>%
  filter(Journal_name %in%
         count_Journal_name$Journal_name) -> count_Journal_name_Year #get counts per
journal and year only for most common journals (n>25)

count_Journal_name_Year %>% ggplot(
      aes(x = str_to_title(Publication.Year),
          y = str_to_title(Journal_name),
          size = n)) +
  geom_point() +
  xlab("Publication year") +
  ylab("Journals with over 20 articles")
```

Articles - journal IF from JCR

Do not run fetching - this data can be loaded from a saved file.

▼ Code

```
## NOTE: Skip this code chunk and use saved data in the next chunk.
# Based on JCR 2022 Impact Factors: https://impactfactorforjournal.com/jcr-impact-factor-
2022/. The website data is not usable - the link changes every year - and only top 100 are
shown.
# Data for 2022 are saved as a pdf: JCR-2021-Impact-Factor-PDF.pdf (password protected for
exports) with IF for 10593 journals.
# Package tabulizer is not longer supported and requires JVR for R, so use pdftools package
instead.
# library(pdftools)

# Extract the table
pdf_IF_table <- pdf_text(here("data", "JCR-2021-Impact-Factor-PDF.pdf"))
#str(pdf_IF_table) #a character vector
#length(pdf_IF_table) #222
#head(pdf_IF_table)
pdf_IF_table2 <- strsplit(pdf_IF_table, "\n") #split on every newline mark
pdf_IF_table2 <- unlist(pdf_IF_table2) #change list into a single vector of stings
#pdf_IF_table2[1:10]
pdf_IF_table2 <- pdf_IF_table2[-c(1:6)] #remove first 6 elements with table headers and
description
pdf_IF_table2 <- trimws(pdf_IF_table2,  which = "left") #remove unnecessary leading white
spaces
pdf_IF_table2 <- str_split(pdf_IF_table2, " {2,}") #split at points with at least two white
spaces
table(lengths(pdf_IF_table2)) #there are few hundreds of misaligned records (1 or 2 strings
only) - need to be removed
pdf_IF_table2 <- pdf_IF_table2[as.logical(lengths(pdf_IF_table2) == 3)] #keep only the list
elements with 3 sub-elements
pdf_IF_df <- do.call(rbind.data.frame, pdf_IF_table2) #convert list to data frame
#str(pdf_IF_df)
names(pdf_IF_df) <- c(
  "Nr",
  "Journal_name",
  "Journal_IF"
)

pdf_IF_df[grepl("[0-9].[0-9]", pdf_IF_df$Journal_name), ] #journal titles with at least 2
numbers
pdf_IF_df <- pdf_IF_df[!grepl("[0-9].[0-9]", pdf_IF_df$Journal_name), ] #remover rows where
journal titles have at least 2 numbers
dim(pdf_IF_df) #10316
readr::write_csv(pdf_IF_df, file = here("data", "pdf_IF_df.csv"))


##Merge and summarise information on IF of journals
#pdf_IF_df <- readr::read_csv(file = here("data", "pdf_IF_df.csv"), show_col_types = FALSE)
#names(pdf_IF_df)
#glimpse(pdf_IF_df) #Journal_name are a mix of title and upper case
pdf_IF_df$Journal_name <- str_to_upper(pdf_IF_df$Journal_name) #change to all capitals
pdf_IF_df$Journal_IF <- as.numeric(pdf_IF_df$Journal_IF) #change to numeric values
#hist(pdf_IF_df$Journal_IF) #strongly right-skewed
#pdf_IF_df[pdf_IF_df$Journal_IF > 100, ]

#check overlap of journal names
length(unique(dat_articles$Journal_name)) #total
length(intersect(pdf_IF_df$Journal_name, dat_articles$Journal_name)) #number  matching
length(setdiff(dat_articles$Journal_name, pdf_IF_df$Journal_name)) #number not matching

#check for & symbol in journal names and do some cleaning in IF data frame before merging
#grep("&", pdf_IF_df$Journal_name, value = TRUE) #lots!
pdf_IF_df$Journal_name <- gsub(" & ", " AND ", pdf_IF_df$Journal_name, fixed = TRUE)
#check overlap of journal names
length(intersect(pdf_IF_df$Journal_name, dat_articles$Journal_name)) #413 matching
length(setdiff(dat_articles$Journal_name, pdf_IF_df$Journal_name)) #340 not matching

#match and merge data to dat_articles by uppercase journal name
dat_articles_IF <- left_join(dat_articles, pdf_IF_df, by = "Journal_name")
#dim(dat_articles_IF)
#str(dat_articles_IF)
#names(dat_articles_IF)
readr::write_csv(dat_articles_IF, file = here("data","dat_articles_IF.csv"))
```

Summarise journal Impact Factor (IF) data.

▼ Code

```
dat_articles_IF <- readr::read_csv(file = here("data","dat_articles_IF.csv"))
#check for missing data on IF
table(is.na(dat_articles_IF$Journal_IF)) #1543 articles have IF, 863 no
```

```
FALSE  TRUE
 1543   863
```

▼ Code

```
#dat_articles_IF %>% drop_na(journal_IF) -> dat_articles_IF #only keep records with IF data
dim(dat_articles_IF)[1] #2406 records
```

```
[1] 2406
```

▼ Code

```
n_distinct(dat_articles_IF$Journal_name) #753 different journals
```

```
[1] 753
```

▼ Code

```
n_distinct(dat_articles_IF$Journal_name[!is.na(dat_articles_IF$Journal_IF)]) #413 different
journals with IF
```

```
[1] 413
```

▼ Code

```
max(na.omit(dat_articles_IF$Journal_IF)) #63.714 highest IF
```

```
[1] 63.714
```

▼ Code

```
#hist(dat_articles_IF$Journal_IF, na.rm = TRUE) #strongly right-skewed distribution
median(dat_articles_IF$Journal_IF, na.rm = TRUE)
```

```
[1] 3.282
```

▼ Code

```
IQR(dat_articles_IF$Journal_IF, na.rm = TRUE)
```

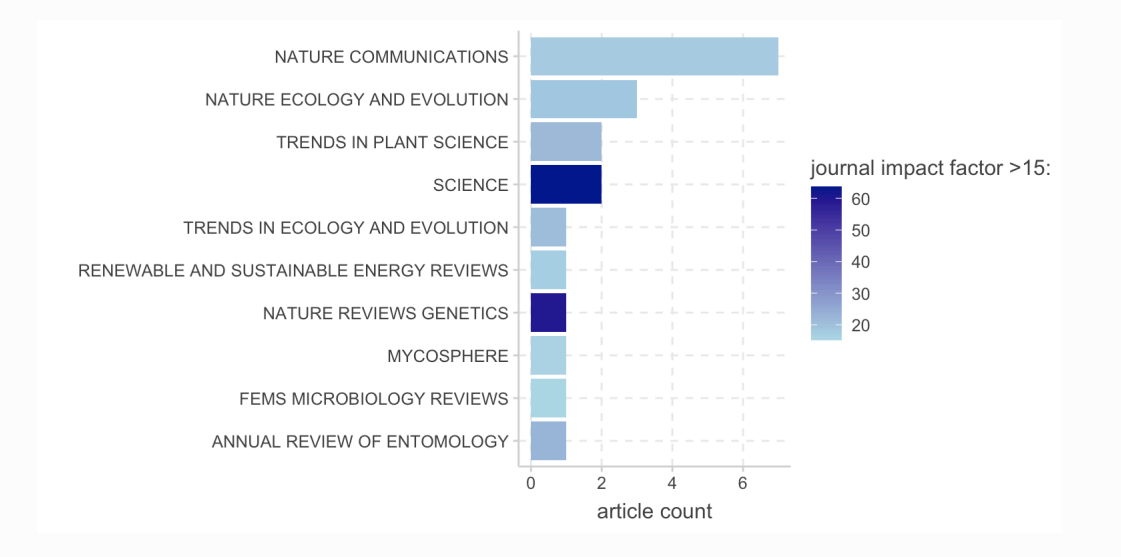
```
[1] 3.225
```

There are 1543 articles in journals with IF, 863 without. Hghest IF is 63.714 (Science, 2 articles).

Plot journals with Impact Factor >15

▼ Code

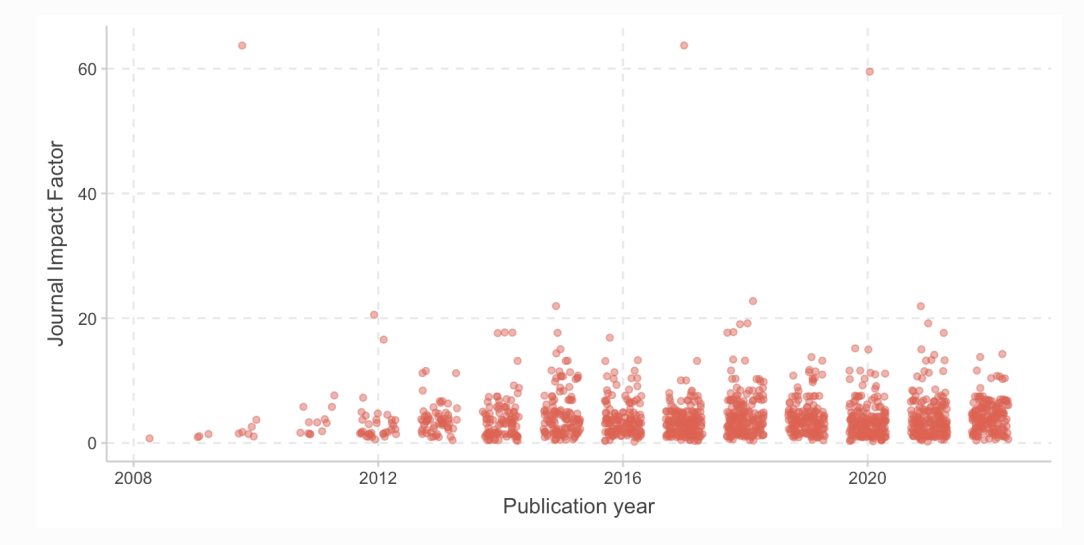
```
dat_articles_IF <- readr::read_csv(file = here("data","dat_articles_IF.csv")) #load saved
fetched data
#get records for journals with IF > 15
dat_articles_IF %>% arrange(desc(Journal_IF)) %>% select(Journal_name, Journal_IF,
Manual.Tags, Title) %>% filter(Journal_IF > 15) -> dat_articles_IF_over15
#dim(dat_articles_IF_over15)[1] #20 in journals with IF >15
#dim(dat_articles_IF_over15)[1]/dim(dat_articles_IF)[1]*100 # 20 is 1% of all articles
dat_articles_IF_over15 %>% add_count(Journal_name, name = "Count") %>% arrange(desc(Count))
-> dat_articles_IF_over15 #counts added as a column to records
#dat_articles_IF_over15$Manual.Tags #see manual tags

#see the journals
dat_articles_IF_over15 %>% distinct(Journal_name, .keep_all = TRUE) %>% arrange(desc(Count))
-> dat_articles_IF_over15_journals #counts as new data frame of journals
#dim(dat_articles_IF_over15_journals)[1] #10 journals with >15 IF

#plot as counts with gradient fill for IF
ggplot(dat_articles_IF_over15_journals, aes(x = reorder(Journal_name, Count), y = Count,
fill = Journal_IF)) + #, fill = "grey"
  geom_bar(stat = "identity") +
  theme(legend.position = "right") +
  coord_flip() +
  xlab("") +
  ylab("article count") +
  scale_fill_gradient(name = 'journal impact factor >15:', low = "lightblue", high =
"darkblue")
```



Plot journal Impact Factor values per publication year.

▼ Code

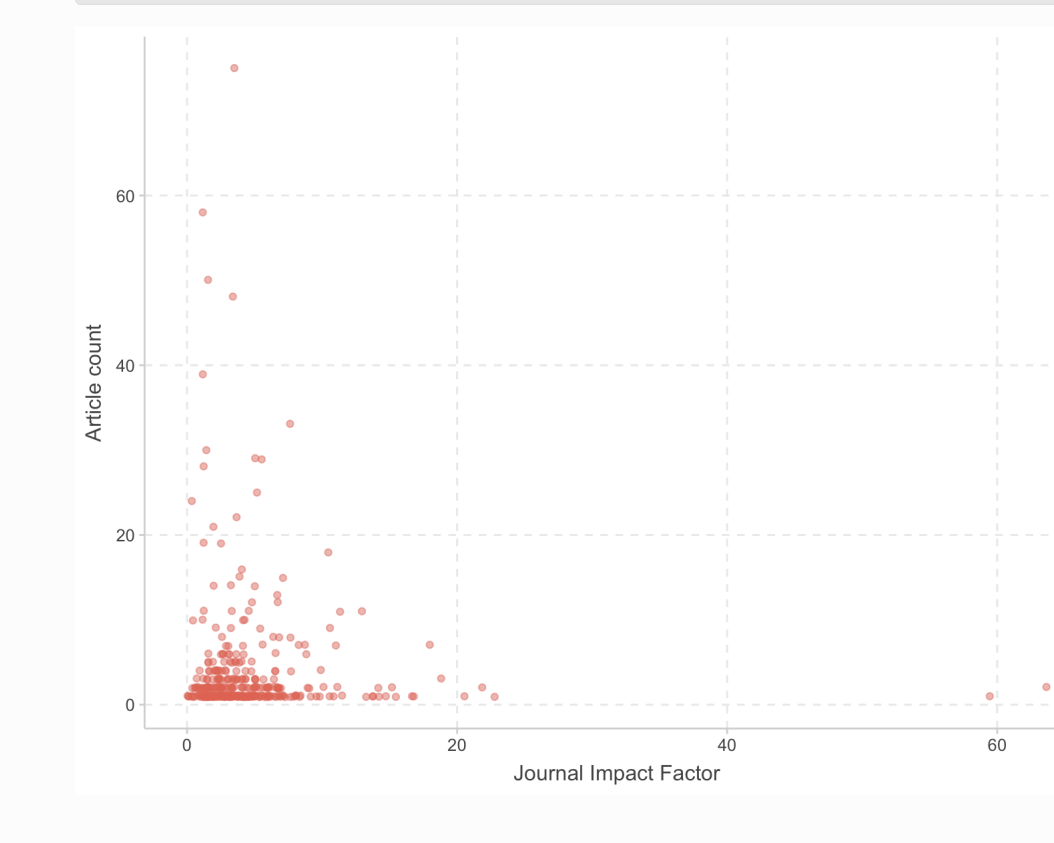
```
#plot number of distinct journal names per year
dat_articles %>%
  filter(!is.na(Journal_IF)) %>%
  #summarise(n = n_distinct(Journal_name)) %>%
  ggplot(
      aes(x = Publication.Year,
          y = Journal_IF)) +
  geom_jitter(position = position_jitter(width = 0.3, height = 0.1), alpha = 0.5) +
  xlab("Publication year") +
  ylab("Journal Impact Factor")
```



Plot journal Impact Factor values versus article counts.

▼ Code

```
#plot article count versus journal impact factor
dat_articles_IF %>%
  filter(!is.na(Journal_IF)) %>%
  count(Journal_IF) %>%
  ggplot(
        aes(x = (Journal_IF),
            y = n)) +
  geom_jitter(position = position_jitter(width = 0.3, height = 0.1), alpha = 0.5) +
  xlab("Journal Impact Factor") +
  ylab("Article count")
```



Articles - Scimago Journal Ranking (SJR)

Using data from Scimago website https://www.scimagojr.com/journalrank.php - downloaded as "scimagojr 2021.csv".
This data is used to get journals Quantile, Rank, Country, Region and Subject Categories.

(Do not run - load saved data object).

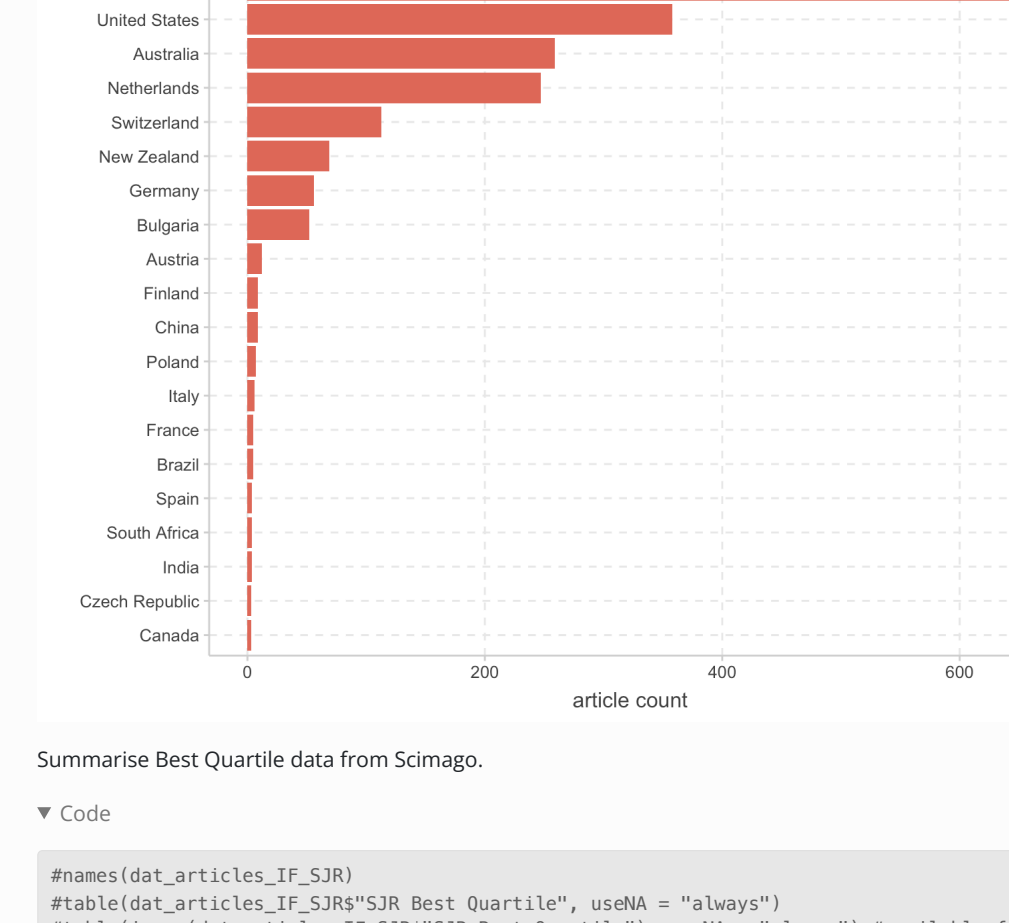
▼ Code

```
dat_scimago <- readr::read_delim(file = here("data","scimagojr 2021.csv"), delim = ";")
attr(dat_scimago,'problems') <- NULL #remove problems attribute from the IF column

#clean-up Scimago journal names:
dat_scimago$Journal_name <- str_replace_all(dat_scimago$Title, "[\r\n]" , " ") %>%
str_to_upper() #remove newline symbols and change all to uppercase
dat_scimago$Journal_name <- gsub("&", "AND", dat_scimago$Journal_name, fixed = TRUE)
#replace all & with AND
#names(dat_scimago)
# glimpse(dat_scimago) #all capitals
# dat_scimago$Title
# dat_scimago$"SJR Best Quartile"
# dat_scimago$Country
# table(dat_scimago$Region)
# dat_scimago$Categories #needs to be strsplit

##Merge SJR data with articles dataframe.

#check the overlap of journal names in two data frames
dat_articles_IF$Journal_name %in% dat_scimago$Journal_name -> fit
sum(fit) #1880 found
sum(!fit) #526 not found
#table(dat_articles_IF$Journal_name[!fit]) #journals not found in Scimago (mostly local
ones, some preprint repos, etc.)

### testing - find "EMU": "EMU" %in% dat_scimago$Journal_name #TRUE
#dat_scimago[dat_scimago$Journal_name=="EMU", ]
#table(str_detect(dat_scimago$Journal_name, "PROCEEDINGS OF THE NATIONAL ACADEMY OF
SCIENCES")) #6 partial matches
#dat_scimago$Journal_name[str_detect(dat_scimago$Journal_name, "PROCEEDINGS OF THE NATIONAL
ACADEMY OF SCIENCES")] #showing matching names - "PROCEEDINGS OF THE NATIONAL ACADEMY OF
SCIENCES OF THE UNITED STATES OF AMERICA" can be fixed
dat_scimago$Journal_name <- str_replace(dat_scimago$Journal_name, "^PROCEEDINGS OF THE
NATIONAL ACADEMY OF SCIENCES OF THE UNITED STATES OF AMERICA$", "PROCEEDINGS OF THE NATIONAL
ACADEMY OF SCIENCES")
#dat_scimago[dat_scimago$Title=="PROCEEDINGS OF THE NATIONAL ACADEMY OF SCIENCES", ]

#remove columns that are not essential
dat_scimago %>% select(-c("Title", "Sourceid", "Type", "Issn", "H index", "Total Docs.
(2021)", "Total Docs. (3years)", "Total Refs.", "Total Cites (3years)", "Citable Docs.
(3years)", "Cites / Doc. (2years)", "Ref. / Doc.", "Publisher", "Coverage")) -> dat_scimago

#merge data frames to append journal IF to my publications table
dat_articles_IF_SJR <- left_join(dat_articles_IF, dat_scimago, by = c("Journal_name"))
dat_articles_IF_SJR$SJR <- as.numeric(gsub(",", ".", dat_articles_IF_SJR$SJR)) #change to
numeric scores
readr::write_csv(dat_articles_IF_SJR, file = here("data","dat_articles_IF_SJR.csv"))
```

Summarise Scimago journal country.

▼ Code

```
dat_articles_IF_SJR <- readr::read_csv(file = here("data","dat_articles_IF_SJR.csv"))
#table(is.na(dat_articles_IF_SJR$SJR)) #1886 available

#counts per journal Country
data.frame(table(dat_articles_IF_SJR$Country)) %>% rename(Country = Var1, Count = Freq) %>%
arrange(-Count) %>% slice_head(n = 20) %>%
ggplot(aes(x = reorder(Country, Count), y = Count)) +
  geom_bar(stat = "identity") +
  coord_flip() +
  xlab("")  +
  ylab("article count")
```



Summarise Best Quartile data from Scimago.

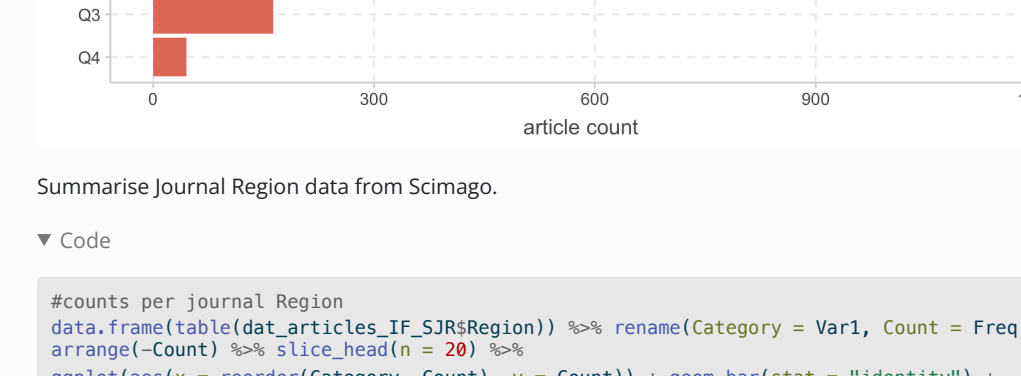
▼ Code

```
#names(dat_articles_IF_SJR)
#table(dat_articles_IF_SJR$"SJR Best Quartile", useNA = "always")
#table(is.na(dat_articles_IF_SJR$"SJR Best Quartile"), useNA = "always") #available for 1886
articles

data.frame(table(dat_articles_IF_SJR$"SJR Best Quartile")) %>% rename(Category = Var1, Count
= Freq) %>% arrange(-Count) %>% filter(Category != "-") %>% filter(Category == "Q1") %>%
select(Count) #number of articles in Q1, 1162 is 62% out of 1886 with quartile info
available
```

	Count
1	1162

▼ Code

```
#counts per best Q category "SJR Best Quartile"
data.frame(table(dat_articles_IF_SJR$"SJR Best Quartile")) %>% rename(Category = Var1, Count
= Freq) %>% arrange(-Count) %>% filter(Category != "-") %>%
ggplot(aes(x = reorder(Category, Count), y = Count)) + geom_bar(stat = "identity") +
  coord_flip() +
  xlab("") +
  ylab("article count")
```



Summarise Journal Region data from Scimago.

▼ Code

```
#counts per journal Region
data.frame(table(dat_articles_IF_SJR$Region)) %>% rename(Category = Var1, Count = Freq) %>%
arrange(-Count) %>% slice_head(n = 20) %>%
ggplot(aes(x = reorder(Category, Count), y = Count)) + geom_bar(stat = "identity") +
  coord_flip() +
  xlab("") +
  ylab("article count")
```

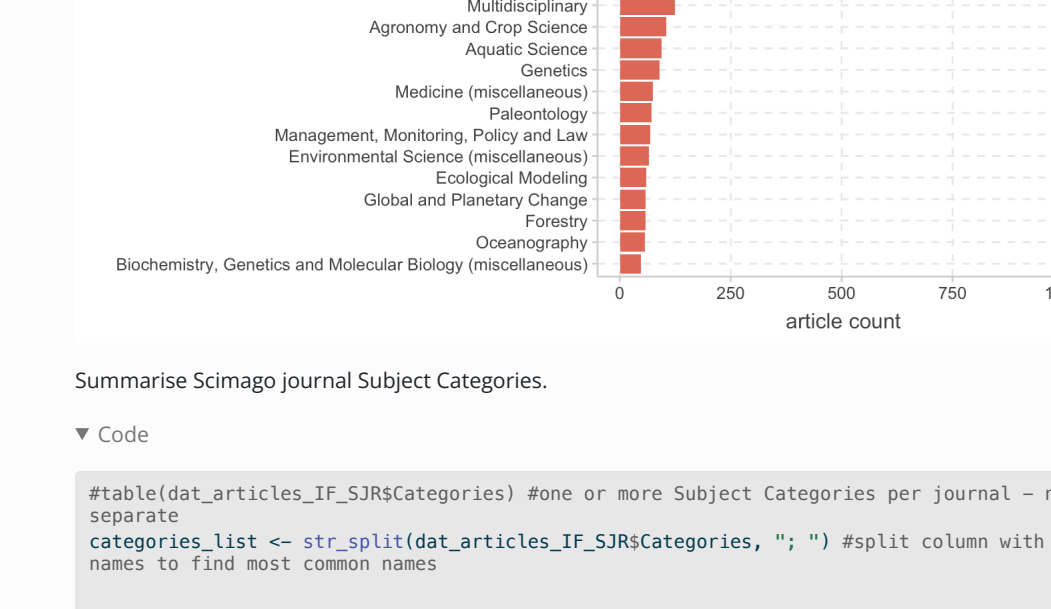
▼ Code

```
# #Mosaic plot for counts per journal Q and Region
# dat_articles_IF_SJR$SJR_Best_Quartile <- dat_articles_IF_SJR$"SJR Best Quartile"
# dat_articles_IF_SJR$SJR_Best_Quartile[dat_articles_IF_SJR$SJR_Best_Quartile == "-"] <- NA
# table(dat_articles_IF_SJR$SJR_Best_Quartile, useNA = "always")
# table(dat_articles_IF_SJR$Region, useNA = "always")
# dat_articles_IF_SJR$Regions <- as.factor(dat_articles_IF_SJR$Region)
# levels(dat_articles_IF_SJR$Regions) <- c("Other", "Other", "Other", "Other", "Other",
"Northern America", "Pacific Region", "Western Europe")
#
# #using package ggmosaic
# dat_articles_IF_SJR %>% filter(!is.na(SJR_Best_Quartile)) %>% ggplot() +
#   geom_mosaic(aes(x = product(Regions, SJR_Best_Quartile), fill = Regions)) +
#   theme_mosaic() +
#   theme(legend.position = "none")
```

Summarise journal category data from Scimago.

▼ Code

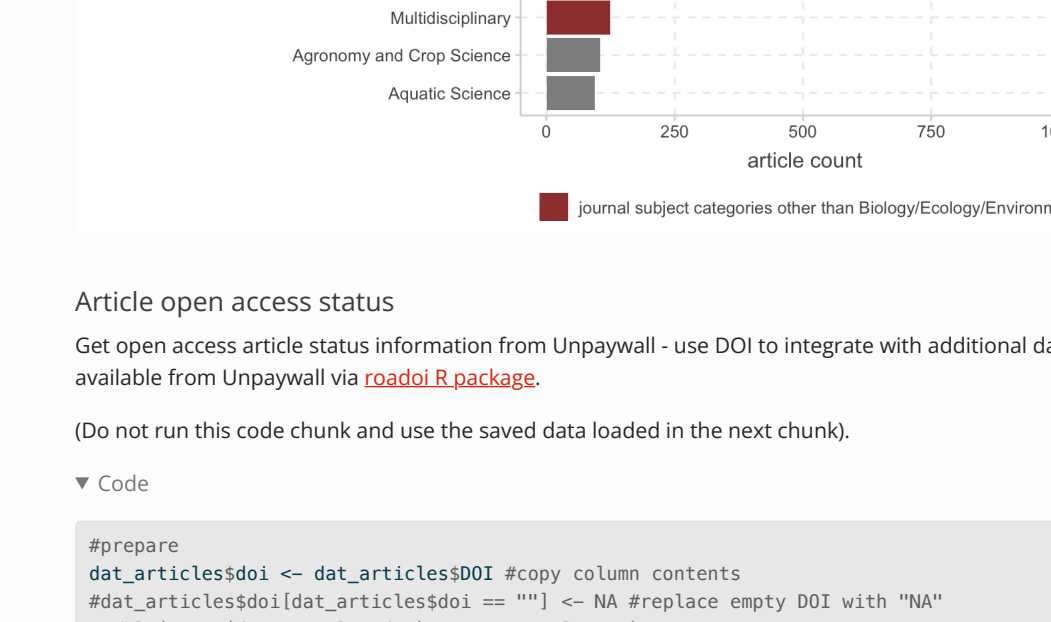
```
#counts by main use type (some articles have more than one, needs cleaning)
dat_articles_jcat <- str_split(dat_articles_IF_SJR$Categories, "; ") %>% unlist() #separate
the categories
dat_articles_jcat <- substring(dat_articles_jcat, 1, nchar(dat_articles_jcat)-5) #remove
last 4 characters
#get top 20 most common journal categories and plot
data.frame(table(dat_articles_jcat)) %>% rename(Category = dat_articles_jcat, Count = Freq)
%>% arrange(-Count) %>% slice_head(n = 20) %>%
ggplot(aes(x = reorder(Category, Count), y = Count)) +
  geom_bar(stat = "identity") +
  coord_flip() +
  xlab("")  +
  ylab("article count")
```



Summarise Scimago journal Subject Categories.

▼ Code

```
#table(dat_articles_IF_SJR$Categories) #one or more Subject Categories per journal - need to
separate
categories_list <- str_split(dat_articles_IF_SJR$Categories, "; ") #split column with author
names to find most common names

##How many categories per journal? - interdisciplinarity
#count number of sub-elements in a list, for each journal
count_categories <- lengths(categories_list) #NA is counted as 1!
count_categories[is.na(categories_list)] <- "no information" #change NA to text value and
all to strings
categories_list_counts <- as.data.frame(table(count_categories)) ##count each value
names(categories_list_counts) <- c("subject_categories", "article_count") #rename variables
levels(categories_list_counts$subject_categories) <- c("1 subject category", "2 subject
categories", "3 subject categories", "4 subject categories", "5 subject categories", "6
subject categories", "7 subject categories", "8 subject categories", "9 subject categories",
"no information") #rename levels

# #plot
# Colors_JCatL <-  c("no information" = "#bdbdbd", rep.int("#2b2b2b", 9) ) #custom colors
# ggplot(categories_list_counts, aes(x = subject_categories, y = article_count, fill =
subject_categories)) +
#   geom_bar(stat = "identity") +
#   coord_flip() +
#   xlab("")  +
#   scale_fill_manual(values = Colors_JCatL) +
#   #geom_text(aes(label = article_count), color="white", size = 3, position =
position_stack(vjust = 0.5)) +
#   theme(legend.position = "none") +
#   scale_x_discrete(limits = rev(levels(categories_list_counts$subject_categories))) +
#reverse order of the bars (factor)
#   ylab("article count")

##Most common journal categories? - dominating research subjects
#unlist and clean categories by removing last 5 characters (Q-values) from each sub-element
categories_unlisted <- gsub('.{5}$', '', unlist(categories_list))
count_categories_unlisted <- as.data.frame(table(categories_unlisted)) %>%
slice_max(order_by = Freq, n = 10) #get top10 most common subject categories

#plot
Colors_JCatUL <-  c("Multidisciplinary" = "#8f2e2e",  "Management, Monitoring, Policy and
Law" = "#8f2e2e", "Medicine (miscellaneous)" = "#8f2e2e", rep.int("#2b2b2b", 17) ) #custom
colors
ggplot(count_categories_unlisted, aes(x = reorder(categories_unlisted, Freq), y = Freq, fill
= categories_unlisted)) +
  geom_bar(stat = "identity") +
  coord_flip() +
  scale_fill_manual(name = '', values = Colors_JCatUL, breaks = c('Multidisciplinary'),
labels = c("journal subject categories other than Biology/Ecology/Environment")) +
    #geom_text(aes(label = article_count), color = "white", size = 3, position =
position_stack(vjust = 0.5)) +
  theme(legend.position = "bottom", legend.direction = "horizontal") +
  xlab("") +
  ylab("article count")
```



Article open access status

Get open access article status information from Unpaywall - use DOI to integrate with additional data available from Unpaywall via roadoi R package.

(Do not run this code chunk and use the saved data loaded in the next chunk).

▼ Code

```
#prepare
dat_articles$doi <- dat_articles$DOI #copy column contents
#dat_articles$doi[dat_articles$doi == ""] <- NA #replace empty DOI with "NA"
#table(is.na(dat_articles$doi), useNA = "always") #2406 not NA, 334 are NA
dois <- unique(dat_articles$doi) #get unique DOI (not empty strings)
#length(dois) # 2073

#connect and fetch the data
OA_df <- roadoi::oadoi_fetch(dois = dois, email = "losialagisz@gmail.com") #get refs data
based on doi
#There were 33 warnings (use warnings() to see them)
readr::write_csv(OA_df, file = here("data","OA_article_data_Unpaywall.csv")) #save fetched
data
```

Merge and summarise information on Open Access status of journal articles from Unpaywall.

▼ Code

```
OA_df <- readr::read_csv(file = here("data","OA_article_data_Unpaywall.csv")) #load saved
data object
# names(OA_df)
# class(OA_df) #list of data frames with lists of data frames
# str(warnings()) #List of warnings - mainly that an article is not in Unpaywall
# failed_DOIs <- warnings() #get failed DOIs
# dim(OA_df) #2040
# length(unique(OA_df$doi)) #2040 - successfully retrieved
# length(anyDuplicated(OA_df$doi)) #1

dat_articles$doi <- dat_articles$DOI #copy column contents for easier merging of the two
dataframes
# length(unique(dat_articles$doi)) #2073
# length(anyDuplicated(dat_articles$doi)) #1

##check overlap
# length(setdiff(dat_articles$doi, OA_df$doi)) #262 not matching (ignores missing doi)
# length(setdiff(OA_df$doi, dat_articles$doi)) #NA
# length(intersect(OA_df$doi, dat_articles$doi)) #1811 matching

##match and merge data to dat_articles by DOI
dat_articles_OA <- left_join(dat_articles, OA_df, by = "doi")
#dim(dat_articles_OA) #2406
#str(dat_articles_OA) #note this also includes author info as nested data frames
#names(dat_articles_OA)

readr::write_csv(dat_articles_OA, file = here("data","dat_articles_OA.csv")) #save merged
data frame for later - this removes nested data frames with author information
```

Processing article open access status information.

▼ Code

```
dat_articles_OA <- readr::read_csv(file = here("data","dat_articles_OA.csv"))
#summarise - note that <NA> is for 595 studies that had doi not matched with info on
Unpaywall, resulting in NA values
# table(!is.na(dat_articles_OA$is_oa)) #data for 1811, missing for 595
# table(dat_articles_OA$is_oa, useNA = "always")
# table(dat_articles_OA$oa_status, useNA = "always")
# table(dat_articles_OA$has_repository_copy, useNA = "always")
# table(dat_articles_OA$journal_is_oa, useNA = "always")
# table(dat_articles_OA$is_oa, dat_articles_OA$oa_status)
```

Open Access status data available for 1811 articles.
Among those, 1096 articles have some type of Open Access version available:
1019 have repository copy,
595 journal is Open Access.

Article citations counts

Get article academic citation data from Crossref using rcrossref package.

(Do not run this code chunk and use saved data in the next chunk).

▼ Code

```
##Using rcrossref package: Register for the polite pool. Open file:
file.edit("~/.Renviron"). Add email address to be shared with Crossref crossref_email=
"name@example.com". Save the file and restart your R session

##Load  a processed version list of "impact" literature related to ALA - use the .csv file
#dat <- read.csv(here("data", "dat_all_processed.csv")) #loading all data
#dat_articles <- dat %>% filter(Item.Type == "journalArticle") #%>%
dplyr::select(Key:Manual.Tags)
# OR
dat_articles <- read.csv(here("data", "dat_articles_processed.csv"))
#dim(dat_articles) #2406
#names(dat_articles)
# dois <- dat$DOI #article doi
# length(unique(dois)) #2171 DOI (but see below)
# dois <- unique(dois)
# dois <- dois[dois != ""]
# length(dois) #2170

#fetch only citation counts
citation_counts <- cr_citation_count(doi = dois, async = TRUE, verbose = TRUE)
#str(citation_counts) #a data frame with just doi and count columns
table(is.na(citation_counts$count)) #2131 ok, 39 NA

#crossref_works <- cr_works(dois = c(dois[1:10])) #using high level API function to get more
information
#str(crossref_works) #messy output format, often no affiliations

#readr::write_csv(citation_counts, file = here("data","citation_counts.csv")) #save fetched
data for later
citation_counts <- readr::read_csv(file = here("data","citation_counts.csv"))

#dat_articles <- readr::read_csv(file = here("data","dat_articles.csv"))
dat_articles$doi <- dat_articles$DOI #copy column contents
dat_articles$doi[dat_articles$doi == ""] <- NA #replace empty DOI with "NA"
table(is.na(dat_articles$doi), useNA = "always") #2740 not NA, 353 are NA
dat_articles_CR <- left_join(dat_articles, citation_counts, by = "doi") #match and merge
data to dat_articles by doi
table(is.na(dat_articles_CR$count)) #2039 counts available

readr::write_csv(dat_articles_CR, file = here("data","dat_articles_CR.csv")) #save merged
data frame for plotting later
```

Summarise and match citation data to the main data frame by doi.

▼ Code

```
dat_articles_CR <- readr::read_csv(file = here("data","dat_articles_CR.csv"))  #load saved
merged data frame
# median(dat_articles_CR$count, na.rm = TRUE) #6
# mean(dat_articles_CR$count, na.rm = TRUE) #19.3
# sd(dat_articles_CR$count, na.rm = TRUE) #96.4
# IQR(dat_articles_CR$count, na.rm = TRUE) #14
# min(dat_articles_CR$count, na.rm = TRUE) #0
# max(dat_articles_CR$count, na.rm = TRUE) #3732

#subset highly cited (>100) publications
dat_articles_CR_highlycited100 <- dat_articles_CR %>%
  filter(count > 100)
#dim(dat_articles_CR_highlycited100) #45 articles with >100 citations

#subset highly cited (>500) publications
dat_articles_CR_highlycited500 <- dat_articles_CR %>%
  filter(count > 500)
dim(dat_articles_CR_highlycited500) #4 articles with >500 citations
```

```
[1]  4 43
```

▼ Code

```
#dat_articles_CR_highlycited[ , 3:6]
```

There are 45, 43 articles with >100 citations.
There are 4, 43 articles with >500 citations.

Plot citation counts across years:

▼ Code

```
pos <- position_jitter(width = 0.1, height = 0.1, seed = 1) #used below to add jitter

#plot
dat_articles_CR %>%
  filter(!is.na(count)) %>% #without articles that have no available citation counts
  ggplot(aes(x = Publication.Year, y = count)) +
  geom_point(alpha = 0.2, position = pos, color = "#66c2a5") +
  #geom_smooth(method = "lm", color = "grey") +
  theme_minimal()  +
  ylab("total citation count") +
  xlab("year") +
  ggrepel::geom_text_repel(data = dat_articles_CR_highlycited500, aes(label =
Publication.Title,), size = 3, max.overlaps = 25, color = "#66c2a5") + #add journal names as
labels for highly cited publications only
  theme(legend.position = "none") #point labels
```

▼ Code

```
#save the figure to pdf
#ggsave(here("plots", "plot.citations_years.pdf"), width = 16, height = 6, units = "cm",
dpi = "retina", scale = 2)
```

Altmetrics

Gather Altmetrics data using a customised function picked up from discontinued rAltmetric R package.
(Do not run this code chunk and use saved data in the next chunk).

▼ Code

```
## NOTE: skip this chunk and use saved query output in the next chunk.
# https://cran.r-project.org/web/packages/rAltmetric/index.html
# Package manual: http://cran.nexr.com/web/packages/rAltmetric/rAltmetric.pdf and example:
https://www.r-project.org/nosvn/pandoc/rAltmetric.html
# NOTE: rAltmetric package was "archived on 2022-06-02 at the maintainer's request".
# API Altmetric info: https://api.altmetric.com/docs/call_fetch.html

#dat_articles <- readr::read_csv(file = here("data","dat_articles.csv")) #or use:
#dat_articles_OA_IF_SJR <- readr::read_csv(file = here("data","dat_articles_OA_IF_SJR.csv"))
#names(dat_articles_OA_IF_SJR)

#define getAltmetrics() function
getAltmetrics <- function(doi = NULL,
                          foptions = list(),
                          ...) {
  if (!is.null(doi)) doi <- stringr::str_c("doi/", doi)
  identifiers <- purrr::compact(list(doi))
  if (!is.null(identifiers)) {
    ids <- identifiers[[1]]
  }
  base_url <- "http://api.altmetric.com/v1/"
  request <- httr::GET(paste0(base_url, ids))
  results <-
    jsonlite::fromJSON(httr::content(request, as = "text"), flatten = TRUE)
  results <- rlist::list.flatten(results)
  class(results) <- "altmetric"
  results
}

#define helper function
altmetric_df <- function(altmetric.object) {
  df <- data.frame(t(unlist(altmetric.object)),
                   stringsAsFactors = FALSE)
}

#run for a single doi:
#Altmetrics.results <- getAltmetrics(doi = "10.3389/fagro.2021.657785") %>% altmetric_df()
#tests for multiple dois
# dois <- list(c("10.3389/fagro.2021.657785", "10.1016/j.biocon.2022.109470"))
# dois <- list(c("10.3316/informit.240264498108007", "10.1089/bio.2021.0127",
"10.1016/j.biocon.2022.109470"))
# dois <- dat_articles_OA_IF_SJR$DOI[!is.na(dat_articles_OA_IF_SJR$DOI)][1:25] @removing
empty doi and subsetting

dois <- dat_articles$DOI #get all dois, or dois <- dat_articles_OA_IF_SJR$DOI #all dois

#run custom function and try to catch errors if not found (articles which have 0 Altmetric
score are not in the database)
Altmetrics.results <- list(dois) %>% purrr::pmap( ~ tryCatch( {getAltmetrics(doi = .x) %>%
altmetric_df()},
                                                    error = function(e) {
                                                      cat(paste0("*NOT_FOUND* ", .x, "\n"))
                                                    }))

length(Altmetrics.results) #2073
#str(Altmetrics.results) #list of data frames of 1 observation and different number of
variables
#table(sapply(Altmetrics.results, is.null)) #554 are NULL - no altmetrics data
#Altmetrics.results[sapply(Altmetrics.results, is.null)] <- NULL #remove empty list elements

#convert list of altmetrics results to cleaner data frame - it drops empty rows
Altmetrics.results_df <- data.table::rbindlist(Altmetrics.results, fill = TRUE)
dim(Altmetrics.results_df) #1519
#str(Altmetrics.results_df)
#names(Altmetrics.results_df) #345 columns - extra author names columns after 110!
#View(Altmetrics.results_df[, 70:125]) #see some columns at the end

readr::write_csv(Altmetrics.results_df, file = here("data","Altmetrics.results_df.csv"))
#save
```

Interpreting Altmetrics key output statistics:
- score - "Altmetric score" = the main Altmetric attention score
- cited_by_policies_count - "Number of policy sources that have mentioned the publication"
- cited_by_msm_count - "Number of the news sources that have mentioned the publication"
- cited_by_feeds_count - "Number of blogs that have mentioned the publication"
- cited_by_tweeters_count - "Number of the twitter accounts that have tweeted this publication"
- cited_by_fbwalls_count - "Number of the pages that have shared on Facebook"
- cited_by_rdts_count - "Number of Reddit threads posted about this publication"
- cited_by_videos_count - "Number of the Youtube/Vimeo channels"
- cited_by_gplus_count - "Number of the accounts that have shared on Google+"

See more on
https://docs.google.com/spreadsheets/d/1ndVY8Q2LOaZO_P_HDmSQulagjeUrS250mAL2N5V8GvY/edit#gid=0

Summarize Altmetrics data.

▼ Code

```
Altmetrics.results_df <- readr::read_csv(file = here("data","Altmetrics.results_df.csv"))
#load saved data file
#plot a histogram of the Altmetric scores (code from https://www.r-
bloggers.com/2015/05/altmetrics-on-citeulike-entries-in-r/):
#hist(as.numeric(Altmetrics.results_df$score), main="", xlab="Altmetric score", breaks=50)
dim(Altmetrics.results_df) #1519 articles with Altmetrics data
```

```
[1] 1519  343
```

▼ Code

```
##names(Altmetrics.results_df)

## Checking and processing to more useful formats
#max(as.numeric(Altmetrics.results_df$score))
#Altmetrics.results_df$authors1 #name of first author (subsequent authors names are stored
in separate numbered author columns)
#table(is.na(Altmetrics.results_df$images.small)) #no NA
#table(Altmetrics.results_df$is_oa, useNA = "always") #all FALSE - not useful
#table(Altmetrics.results_df$subjects, useNA = "always") #with NA - not useful
#Altmetrics.results_df$history.1y
#table(is.na(Altmetrics.results_df$score)) #no NA - overall composite altmetric score
Altmetrics.results_df$score <- as.numeric(Altmetrics.results_df$score) #make numeric
#Altmetrics.results_df$published_on #in POSIXct format - the number of seconds from 1
January 1970 (or 1960?)
Altmetrics.results_df$published_date <- Altmetrics.results_df$published_on #add a new column
Altmetrics.results_df$published_date <- as.numeric(gsub("NA           ", NA,
Altmetrics.results_df$published_date)) #change "NA          " to NA and convert to numeric
Altmetrics.results_df$published_date <- as.POSIXct(Altmetrics.results_df$published_date,
origin = "1970-01-01") #convert POSIXct
Altmetrics.results_df$year <- as.numeric(format(Altmetrics.results_df$published_date,
format="%Y")) #extract year
#table(is.na(Altmetrics.results_df$year)) #4 values missing
```

There are 1519 articles with Altmetrics data.

▼ Code

```
### Make tables of article subsets with specific types of attention.
##Patents
table(Altmetrics.results_df$cited_by_patents_count, useNA = "always") # 3 articles cited
once each, 1 more than once
```

```
   1    3 <NA>
   3    1 1515
```

▼ Code

```
##table of articles with patent citations
Altmetrics.results_df %>% filter(as.numeric(cited_by_patents_count) > 0) %>% select(title,
journal, authors1, year, doi, cited_by_patents_count) %>%
arrange(desc(cited_by_patents_count)) %>% DT::datatable(rownames = FALSE, width = "100%",
options = list(dom = 't', scrollY = '800px', pageLength = 20), caption = "Table of
Altmetrics for articles cited by patents")
```

Table of Altmetrics for articles cited by patents

title	journal	authors1	year	doi	cited_by_patents_
The Antibacterial Activity of Australian Leptospermum Honey Correlates with Methylglyoxal Levels	PLOS ONE	Nural N. Cokcetin	2016	10.1371/journal.pone.0167780	
Capturing Natural-Colour 3D Models of Insects for Species Discovery and Diagnostics	PLOS ONE	Chuong V. Nguyen	2014	10.1371/journal.pone.0094346	
Toward an ecoregion scale evaluation of eDNA metabarcoding primers: A case study for the freshwater fish biodiversity of the Murray-Darling Basin (Australia)	Ecology and Evolution	Jonas Bylemans	2018	10.1002/ece3.4387	
Plant salt tolerance:	Annals of	Timothy J.			

▼ Code

```
##Policies
table(Altmetrics.results_df$cited_by_policies_count, useNA = "always") # 66 articles cited
once each, 10 more than once
```

```
   1    2    3 <NA>
  66    7    3 1443
```

▼ Code

```
##table of articles with patent citations
Altmetrics.results_df %>% filter(as.numeric(cited_by_policies_count) > 0) %>% select(title,
journal, authors1, year, doi, cited_by_policies_count) %>%
arrange(desc(cited_by_policies_count))  %>% DT::datatable(rownames = FALSE, width = "100%",
options = list(dom = 't', scrollY = '800px', pageLength = 20), caption = "Table of
Altmetrics for articles cited by patents")
```

Table of Altmetrics for articles cited by patents

title	journal	authors1	year	doi	cited_
Climate change influences on global distributions of dengue and chikungunya virus vectors	Philosophical Transactions of the Royal Society B: Biological Sciences	Lindsay P. Campbell	2015	10.1098/rstb.2014.0135	
A statistical explanation of MaxEnt for ecologists	Diversity & Distributions	Jane Elith	2010	10.1111/j.1472-4642.2010.00725.x	
Overview of distribution patterns of zooxanthellate Scleractinia	Frontiers in Marine Science	John Veron	2015	10.3389/fmars.2014.00081	
A decision tree for assessing the risks and benefits of publishing biodiversity data	Nature Ecology & Evolution	Ayesha I. T. Tulloch	2018	10.1038/s41559-018-0608-1	
Dealing with Cumulative Biodiversity Impacts in Strategic Environmental Assessment: A New Frontier for Conservation Planning	Conservation Letters	Amy L. Whitehead	2016	10.1111/conl.12260	
Opening the black box: an open-source	Ecography	Steven J. Phillips	2017	10.1111/ecog.03049	

▼ Code

```
##Twitter (some high numbers here)
table(Altmetrics.results_df$cited_by_tweeters_count, useNA = "always")
```

```
   1    2    3    4    5    6    7    8    9   10   11   12   13   14   15   16
 208  132  101   90   68   40   47   40   41   42   31   22   32   30   25   18
  17   18   19   20   21   22   23   24   25   26   27   28   29   30   31   32
  20   19   19   14   17   11    8   16   11    8   10   14   12   10    8    7
  33   34   35   36   37   38   39   40   41   42   43   44   45   46   47   48
  11    6    3    3    4   10    5    4   10    2    5    2    7    2    4    6
  49   50   51   52   53   55   56   57   58   59   62   63   64   65   66   67
   7    3    3    3    6    5    6    4    4    2    3    2    2    1    2    1
  68   69   71   72   73   74   75   76   77   79   80   82   83   85   86   87
   4    1    2    3    2    3    1    2    1    1    1    2    1    3    1    3
  90   93   95   98  100  101  102  103  107  111  112  113  114  117  123  124
   1    1    2    1    1    1    1    1    1    2    1    1    1    2    1    1
 131  133  134  137  141  146  148  150  152  156  174  181  186  190  198  204
   1    1    2    1    1    1    1    1    1    1    1    1    1    1    1    1
 219  243  276  322  323  364  390  453  598  618 1456 <NA>
   1    1    1    1    1    1    1    1    1    1    1  132
```

▼ Code

```
sum(table(Altmetrics.results_df$cited_by_tweeters_count)) # 1387 articles mentioned at least
once
```

```
[1] 1387
```

▼ Code

```
sum(Altmetrics.results_df$cited_by_tweeters_count, na.rm = TRUE) #28730 total numbers of
mentions
```

```
[1] 28730
```

▼ Code

```
##table of articles with cited_by_tweeters_count >500
Altmetrics.results_df %>% filter(as.numeric(cited_by_tweeters_count) > 100) %>%
select(title, journal, authors1, year, doi, cited_by_tweeters_count) %>%
arrange(desc(cited_by_tweeters_count)) %>% DT::datatable(rownames = FALSE, width = "100%",
options = list(dom = 't', scrollY = '800px', pageLength = 20), caption = "Table of
Altmetrics for articles with Tweeter counts > 100")
```

Table of Altmetrics for articles with Tweeter counts > 100

title	journal	authors1	year	doi	cited_
The number of tree species on Earth	Proceedings of the National Academy of Sciences of the United States of America	Roberto Cazzolla Gatti	2022	10.1073/pnas.2115329119	
Birds, butterflies and flowers in the tropics are not more colourful than those at higher latitudes	Global Ecology & Biogeography	Rhiannon L. Dalrymple	2015	10.1111/geb.12368	
Striped bodypainting protects against horseflies	Royal Society Open Science	Gábor Horváth	2019	10.1098/rsos.181325	
A new species of bandy-bandy (Vermicella: Serpentes: Elapidae) from the Weipa region, Cape York, Australia	ZOOTAXA	CHANTELLE M. DEREZ	2018	10.11646/zootaxa.4446.1.1	
Global Patterns and Drivers of Bee Distribution	Current Biology	Michael C. Orr	2021	10.1016/j.cub.2020.10.053	
Loss of vocal culture and	Proceedings of				

▼ Code

```
##Facebook
table(Altmetrics.results_df$cited_by_fbwalls_count, useNA = "always")
```

```
   1    2    3    4    5    6    7    8    9   10   11   14   16 <NA>
 276   96   30   22   11    3    5    3    1    2    1    1    1 1067
```

▼ Code

```
sum(table(Altmetrics.results_df$cited_by_fbwalls_count)) #452 articles mentioned at least
once
```

```
[1] 452
```

▼ Code

```
sum(Altmetrics.results_df$cited_by_fbwalls_count, na.rm = TRUE) #848 total numbers of
mentions
```

```
[1] 848
```

▼ Code

```
##table of articles with cited_by_fbwalls_count >10
Altmetrics.results_df %>% filter(as.numeric(cited_by_fbwalls_count) > 10) %>% select(title,
journal, authors1, year, doi, cited_by_fbwalls_count) %>%
arrange(desc(cited_by_fbwalls_count)) %>% DT::datatable(rownames = FALSE, width = "100%",
options = list(dom = 't', scrollY = '800px', pageLength = 20), caption = "Table of
Altmetrics for articles with Facebook counts > 10")
```

Table of Altmetrics for articles with Facebook counts > 10

title	journal	authors1	year	doi	cited_by_
GlobalTreeSearch: The first complete global database of tree species and country distributions	Journal of Sustainable Forestry	E. Beech	2017	10.1080/10549811.2017.1310049	
Introducing the Global Register of Introduced and Invasive Species	Scientific Data	Shyama Pagad	2018	10.1038/sdata.2017.202	
Contribution of citizen science towards international biodiversity monitoring	Biological Conservation	Mark Chandler	2017	10.1016/j.biocon.2016.09.004	

▼ Code

```
##Number of blogs that have mentioned the publication
table(Altmetrics.results_df$cited_by_feeds_count, useNA = "always")
```

```
   1    2    3    4    5    6    7    8    9   11   12   16   17 <NA>
 223   69   17   11   11    6    1    1    1    4    1    1    1 1172
```

▼ Code

```
sum(table(Altmetrics.results_df$cited_by_feeds_count)) #347 mentioned at least once
```

```
[1] 347
```

▼ Code

```
sum(Altmetrics.results_df$cited_by_feeds_count, na.rm = TRUE) #660 total numbers of mentions
```

```
[1] 660
```

▼ Code

```
##table of articles with cited_by_feeds_count >10
Altmetrics.results_df %>% filter(as.numeric(cited_by_feeds_count) > 10) %>% select(title,
journal, authors1, year, doi, cited_by_feeds_count) %>% arrange(desc(cited_by_feeds_count))
%>% DT::datatable(rownames = FALSE, width = "100%", options = list(dom = 't', scrollY =
'800px', pageLength = 20), caption = "Table of Altmetrics for articles with blog feeds
counts > 10")
```

Table of Altmetrics for articles with blog feeds counts > 10

title	journal	authors1	year	doi	cited_by_feed
The number of tree species on Earth	Proceedings of the National Academy of Sciences of the United States of America	Roberto Cazzolla Gatti	2022	10.1073/pnas.2115329119	
Global Patterns and Drivers of Bee Distribution	Current Biology	Michael C. Orr	2021	10.1016/j.cub.2020.10.053	
Urbanisation at Multiple Scales Is Associated with Larger Size and Higher Fecundity of an Orb-Weaving Spider	PLOS ONE	Elizabeth C. Lowe	2014	10.1371/journal.pone.0105480	
Striped bodypainting protects against horseflies	Royal Society Open Science	Gábor Horváth	2019	10.1098/rsos.181325	
A stitch in time – Synergistic impacts to platypus metapopulation extinction risk	Biological Conservation	Gilad Bino	2020	10.1016/j.biocon.2019.108399	

▼ Code

```
##News sources that have mentioned the publication
table(Altmetrics.results_df$cited_by_msm_count, useNA = "always")
```

```
   1    2    3    4    5    6    7    8    9   10   11   12   13   14   15   16
  76   36   25   21   21   22   10    9    5    7    7    8    2    4    1    3
  17   18   21   22   23   24   26   27   29   39   40   42   43   46   47   58
   1    1    3    1    1    2    1    2    2    1    1    2    1    1    1    1
  74   91   92   99  111  113  126  130  164  177  261 <NA>
   1    2    1    1    1    1    1    1    1    1    1 1228
```

▼ Code

```
sum(table(Altmetrics.results_df$cited_by_msm_count)) #291 mentioned at least once
```

```
[1] 291
```

▼ Code

```
sum(Altmetrics.results_df$cited_by_msm_count, na.rm = TRUE) #3334 total numbers of mentions
```

```
[1] 3334
```

▼ Code

```
##table of articles with cited_by_msm_count >100
Altmetrics.results_df %>% filter(as.numeric(cited_by_msm_count) > 100) %>% select(title,
journal, authors1, year, doi, cited_by_msm_count) %>% arrange(desc(cited_by_msm_count)) %>%
DT::datatable(rownames = FALSE, width = "100%", options = list(dom = 't', scrollY = '800px',
pageLength = 20), caption = "Table of Altmetrics for articles with news sources counts >
10")
```

Table of Altmetrics for articles with news sources counts > 10

title	journal	authors1	year	doi	cited_by_msm_
The number of tree species on Earth	Proceedings of the National Academy of Sciences of the United States of America	Roberto Cazzolla Gatti	2022	10.1073/pnas.2115329119	
Loss of vocal culture and fitness costs in a critically endangered songbird	Proceedings of the Royal Society B: Biological Sciences	Ross Crates	2021	10.1098/rspb.2021.0225	
Museum genomics reveals the rapid decline and extinction of Australian rodents since European settlement	Proceedings of the National Academy of Sciences of the United States of America	Emily Roycroft	2021	10.1073/pnas.2021390118	
The platypus: evolutionary history, biology, and an uncertain future	Journal of Mammalogy	Gilad Bino	2019	10.1093/jmammal/gyz058	
Evidence for a complex evolutionary	Biological	Perry G.			

▼ Code

```
##Wikipedia
table(Altmetrics.results_df$cited_by_wikipedia_count, useNA = "always")
```

```
   1    2    3    4    5    6    7    8    9   10   11   12   14   19   41   77
 164   52   22   13    9    4    1    2    2    2    2    1    1    1    1    1
  96  131 <NA>
   1    1 1239
```

▼ Code

```
sum(table(Altmetrics.results_df$cited_by_wikipedia_count)) #280 cited at least once
```

```
[1] 280
```

▼ Code

```
sum(Altmetrics.results_df$cited_by_wikipedia_count, na.rm = TRUE) #928 total numbers of
mentions
```

```
[1] 928
```

▼ Code

```
##table of articles with cited_by_wikipedia_count >10
Altmetrics.results_df %>% filter(as.numeric(cited_by_wikipedia_count) > 10) %>%
select(title, journal, authors1, year, doi, cited_by_wikipedia_count) %>%
arrange(desc(cited_by_wikipedia_count)) %>% DT::datatable(rownames = FALSE, width = "100%",
options = list(dom = 't', scrollY = '800px', pageLength = 20), caption = "Table of
Altmetrics for articles with Wikipedia counts > 10")
```

Table of Altmetrics for articles with Wikipedia counts > 10

title	journal	authors1	year	doi	cit
A monograph of Daviesia (Mirbelieae, Faboideae, Fabaceae)	Phytotaxa	MICHAEL D. CRISP	2017	10.11646/phytotaxa.300.1.1	
Muelleria	Biodiversity Heritage Library	Royal Botanic Gardens (Vic.)	1989	10.5962/bhl.title.112965	
Systematic dismantlement ofLichenostomusimproves the basis for understanding relationships within the honeyeaters (Meliphagidae) and the historical development of Australo-Papuan bird communities	Emu	Árpád S. Nyári	2016	10.1071/mu10047	
The Global Urban Tree Inventory: A database of the diverse tree flora that inhabits the world's cities	Global Ecology & Biogeography	Alessandro Ossola	2020	10.1111/geb.13169	
Global Diversity of Sponges (Porifera)	PLOS ONE	Rob W. M. Van Soest	2012	10.1371/journal.pone.0035105	
Deep-reef fish assemblages of the Great Barrier Reef shelf-break (Australia)	Scientific Reports	Tiffany L. Sih	2017	10.1038/s41598-017-11452-1	
	Proceedings of the				

▼ Code

```
##The sum of all "cited_by" entries (profiles per data source)
table(Altmetrics.results_df$cited_by_accounts_count, useNA = "always") #across all above+
```

```
   1    2    3    4    5    6    7    8    9   10   11   12   13   14   15   16
 229  140   98   98   65   56   52   47   42   38   30   32   35   29   24   18
  17   18   19   20   21   22   23   24   25   26   27   28   29   30   31   32
  26   23   19   18   18   14   11   12   17   12    8   15   10   12    7   10
  33   34   35   36   37   38   39   40   41   42   43   44   45   46   47   48
   7    6    4    3    8    4    7    6    6    4    3    5    7    4    2    7
  49   50   51   52   53   54   55   56   57   58   59   60   61   62   63   64
   7    5    5    2    6    5    4    3    6    6    3    5    1    1    4    1
  65   66   67   68   69   70   71   72   73   74   75   76   77   78   81   84
   1    3    1    2    2    3    2    1    3    2    3    1    5    3    1    1
  85   86   87   89   90   91   93   94   96   98   99  100  102  103  108  109
   1    2    1    1    2    1    1    1    1    1    2    1    1    2    1    2
 110  111  112  113  114  118  119  120  121  123  126  129  130  131  134  141
   1    1    3    1    1    1    1    2    1    1    1    2    1    2    1    1
 143  144  145  156  157  158  159  163  165  168  170  179  198  200  202  211
   1    1    1    1    1    1    1    1    1    2    1    1    1    1    1    1
 224  227  264  280  301  338  431  491  506  551  628  663 1752 <NA>
   1    1    1    1    1    2    1    1    1    1    1    1    1    0
```

▼ Code

```
##table of articles with cited_by_accounts_count >10
Altmetrics.results_df %>% filter(as.numeric(cited_by_accounts_count) > 100) %>%
select(title, journal, authors1, year, doi, cited_by_accounts_count) %>%
arrange(desc(cited_by_accounts_count)) %>% DT::datatable(rownames = FALSE, width = "100%",
options = list(dom = 't', scrollY = '800px', pageLength = 20), caption = "Table of
Altmetrics for articles with all mentions counts > 10")
```

Table of Altmetrics for articles with all mentions counts > 10

title	journal	authors1	year	doi	cited_by
The number of tree species on Earth	Proceedings of the National Academy of Sciences of the United States of America	Roberto Cazzolla Gatti	2022	10.1073/pnas.2115329119	
Striped bodypainting protects against horseflies	Royal Society Open Science	Gábor Horváth	2019	10.1098/rsos.181325	
Birds, butterflies and flowers in the tropics are not more colourful than those at higher latitudes	Global Ecology & Biogeography	Rhiannon L. Dalrymple	2015	10.1111/geb.12368	
Loss of vocal culture and fitness costs in a critically endangered songbird	Proceedings of the Royal Society B: Biological Sciences	Ross Crates	2021	10.1098/rspb.2021.0225	
A new species of bandy-bandy (Vermicella: Serpentes: Elapidae) from the Weipa region,	ZOOTAXA	CHANTELLE M. DEREZ	2018	10.11646/zootaxa.4446.1.1	

▼ Code

```
#hist(as.numeric(Altmetrics.results_df$context.journal.pct)) #which was in top 95% of the
journal
#View(Altmetrics.results_df[Altmetrics.results_df$context.journal.pct >= 95, 1:20] )
#highest score per given journal
```

Match altmetrics data to main data.

▼ Code

```
#match and merge data to dat_articles by DOI
dat_articles_A <- left_join(dat_articles, Altmetrics.results_df, by = "doi")
readr::write_csv(dat_articles_A, file = here("data","dat_articles_A.csv")) #save merged data
frame for plotting later
#dat_articles_A <- readr::read_csv(file = here("data","dat_articles_A.csv"))
```

Find articles with highest Altmetric scores.

▼ Code

```
dat_articles_A <- readr::read_csv(file = here("data","dat_articles_A.csv")) #load processed
and merged Altmetric data
Altmetrics.results_df <- dat_articles_A

#subset highly scoring (>100) publications
Altmetrics.results_df_highlyscored100 <- Altmetrics.results_df %>%
  filter(score > 100)
dim(Altmetrics.results_df_highlyscored100) #79 articles with >100 score
```

```
[1]  76 386
```

▼ Code

```
#subset highly scoring (>1000) publications
Altmetrics.results_df_highlyscored1000 <- Altmetrics.results_df %>%
  filter(score > 1000)
dim(Altmetrics.results_df_highlyscored1000) #4 articles with >1000 score
```

```
[1]   4 386
```

▼ Code

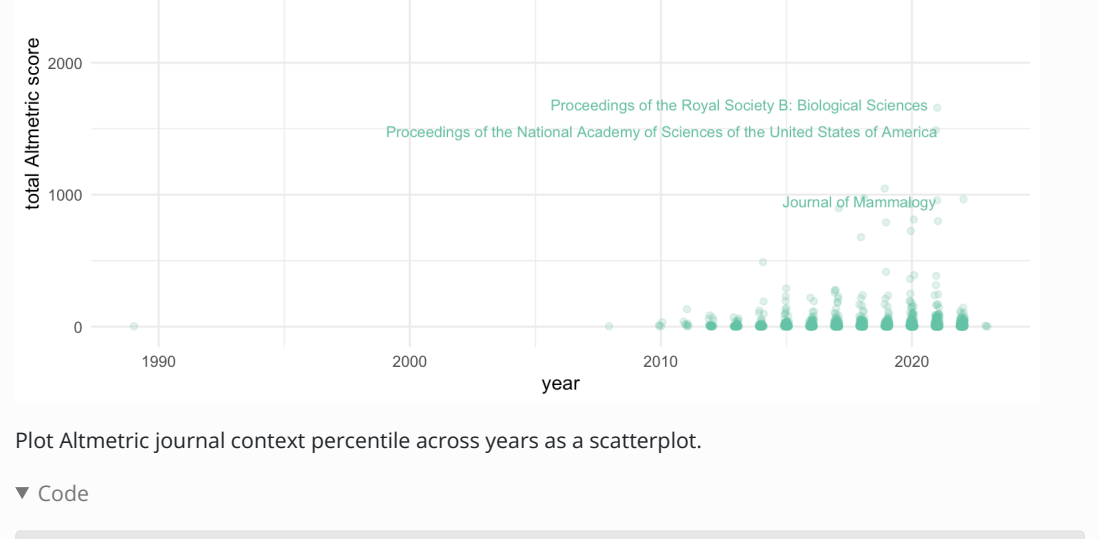
```
#Altmetrics.results_df_highlyscored$journal
#Altmetrics.results_df_highlyscored$authors1
```

There are 76 articles with >100 Altmetric score.
There are 4 articles with >1000 Altmetric score.

Plot Altmetric scores across article publication years:

▼ Code

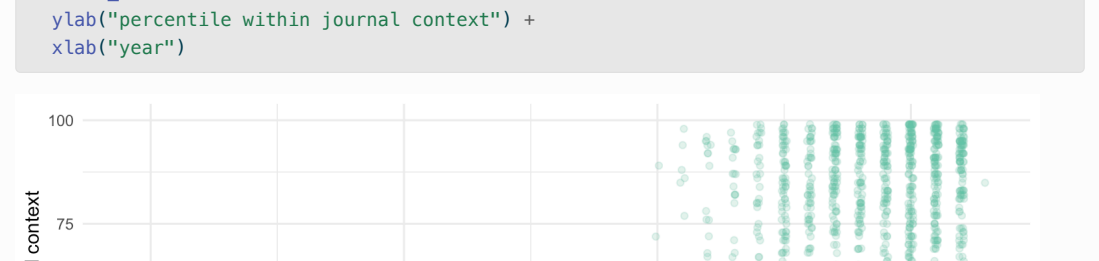
```
pos <- position_jitter(width = 0.1, height = 0.1, seed = 1) #used below to add jitter

#plot
Altmetrics.results_df %>%
  filter(!is.na(score)) %>% #without articles that have no scores
  filter(!is.na(year)) %>% #without articles that have no year
  ggplot(aes(x = year, y = score)) +
  geom_point(color = "#66c2a5", alpha = 0.2, position = pos) +
  #geom_smooth(method = "lm", color = "grey") +
  theme_minimal()  +
  ylab("total Altmetric score") +
  xlab("year") +
  ggrepel::geom_text_repel(data = Altmetrics.results_df_highlyscored1000, aes(label =
journal), size = 3, max.overlaps = 20, color = "#66c2a5") + theme(legend.position = "none")
#point labels
```

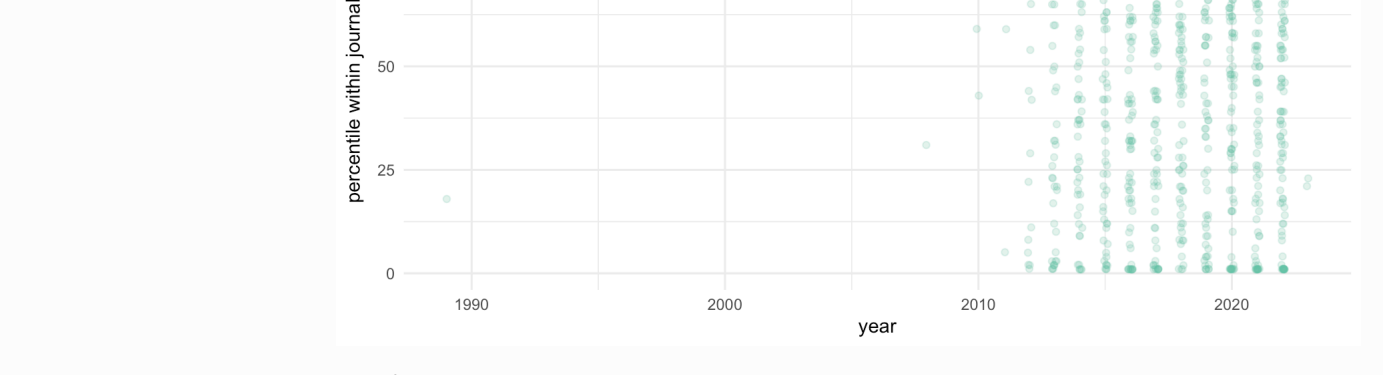


Plot Altmetric journal context percentile across years as a scatterplot.

▼ Code

```
#Altmetrics.results_df$context.journal.pct
Altmetrics.results_df %>%
  filter(!is.na(year)) %>% #without articles that have no year
  ggplot(aes(x = year, y = as.numeric(context.journal.pct))) +
  geom_point(alpha = 0.2, position = pos, color = "#66c2a5") +
  #geom_smooth(method = "lm", color = "grey") +
  theme_minimal() +
  ylab("percentile within journal context") +
  xlab("year")
```





▼ Code

```
# ggrepel::geom_text_repel(aes(label = journal, color = "grey"), size = 3, max.overlaps =
20) + theme(legend.position = "none") #point labels
```

Final plots

The final set of plots for the main manuscript text - the plots themselves are not part of the Supplementary Information, so the code is not automatically evaluated and the figures are not displayed in the html document.

All plots here can be saved as files separate pdf files. Each plot is created primarily from the saved pre-processed data files, so no fetching from other remote sources or external data files is required. Thus, the code in this section can be run independently of all the earlier sections. Data wrangling is only done to produce the final plots in Figures 2 - 9 for the manuscript and data for Figure 1.

Figure 1 - counts of records in Living Atlases.

Using package galah to gather data used in panel B of the figure.

▼ Code

```
##get an overview of the Living Atlases
atlas_counts()

show_all(atlases) #11
atlases <- show_all(atlases)
#atlases$region

#galah_config(atlas = "ALA")  #just for Australia

## print out cutrrent record counts for each Living Atlas, by region
for(i in 1:length(atlases$region)) {
  galah_config(atlas = atlases$region[i])
  out <- paste0(atlases$region[i], " = ", atlas_counts())  # Some output
  print(out)                                               # Using print function
}

#citation("galah")
```

Figure 2 - item types counts across publication years.

Plot year and literature type.

▼ Code

```
# Load processed version list of "impact" literature related to ALA - use the .csv file.
dat_all <- read.csv(here("data", "dat_all_processed.csv"))
#dim(dat_all) #3764   96

dat_all %>%
  count(Publication.Year, Item.Type) -> count_year

nb.cols <- length(unique(dat_all$Item.Type))
mycolors <- colorRampPalette(brewer.pal(8, "Set2"))(nb.cols)

figure2 <- ggplot(count_year, aes(x = as.character(Publication.Year), y = n, fill =
reorder(Item.Type, n))) +
  geom_bar(stat = "identity") + #, position = "dodge"
  xlab("") +
  scale_fill_manual(values = rev(mycolors), name = 'ALA-linked output type:') +
  ylab("item count")  +
  theme(legend.position = "bottom")

#save the figure to pdf
ggsave(plot = figure2, here("plots", "plot.type_years_counts.pdf"), width = 16, height = 12,
units = "cm",  dpi = "retina", scale = 1.5)
```

Figure 3 - comparing scopes of articles vs grey publications.

Panels with article scope categories using manual tags.

▼ Code

```
dat_all <- read.csv(here("data", "dat_all_processed.csv"))
dim(dat_all) #3764   96
#Subset data into peer-reviewed vs. grey literature.
dat_all_articles <- dat_all %>% filter(Item.Type == "journalArticle") %>%
dplyr::select(Key:Manual.Tags)
dim(dat_all_articles) #2406
dat_all_grey <- dat_all %>% filter(Item.Type != "journalArticle") %>%
dplyr::select(Key:Manual.Tags)
dim(dat_all_grey) #1358

# Plot as counts with Custom Color Palettes (color-blind friendly) and labels
Colors_TF <-  c("TRUE" = "#E69F00", "FALSE" = "#56B4E9")
Colors_7light <- c('#99DDFF', '#AAAA00', '#EE8866', '#EEDD88', '#FFAABB', '#44BB99',
'#77AADD')
Colors_6bright <- c('#EE6677', '#228833', '#4477AA', '#CCBB44', '#66CCEE', '#AA3377')

##  Main use of ALA - based on manual tags - by publication type

#for articles
mtags_articles <- str_split(dat_all_articles$Manual.Tags, "; ")
mtags_articles_num <- unlist(mtags_articles)[str_detect(unlist(mtags_articles), "^[0-9]")]
#keep only the 6 tags starting with numerics (main use)
count_mtags_articles_num <- as.data.frame(table(mtags_articles_num)) #739 used
count_mtags_articles_num <- rename(count_mtags_articles_num, Tag = mtags_articles_num)
count_mtags_articles_num$Subset <- "articles"

##for grey
mtags_grey <- str_split(dat_all_grey$Manual.Tags, "; ")
mtags_grey_num <- unlist(mtags_grey)[str_detect(unlist(mtags_grey), "^[0-9]")] #keep only
the 6 tags starting with numerics (main use)
count_mtags_grey_num <- as.data.frame(table(mtags_grey_num)) #only 181 used ALA resources
count_mtags_grey_num <- rename(count_mtags_grey_num, Tag = mtags_grey_num)
count_mtags_grey_num$Subset <- "grey"

#merge into a single dataframe
count_mtags_df <- rbind(count_mtags_grey_num, count_mtags_articles_num)
count_mtags_df$Subset <- factor(count_mtags_df$Subset, levels = c("grey","articles"))
#reorder levels for plotting
count_mtags_df$Tag <- factor(count_mtags_df$Tag, levels = c("6 - ALA published", "4 - ALA
acknowledged", "3 - ALA discussed", "5 - ALA mentioned", "2 - ALA cited", "1 - ALA used"))

#plot
figure3A <- ggplot(count_mtags_df, aes(x = Subset, y = Freq, fill = Tag)) +
  geom_bar(position = "fill", stat = "identity") +
  coord_flip() +
  xlab("") +
  ylab("proportion") +
  #scale_x_discrete(labels = c("articles", "grey")) +
  scale_fill_manual(name = "ALA use type", values = Colors_6bright, breaks = c("1 - ALA
used", "2 - ALA cited", "3 - ALA discussed", "4 - ALA acknowledged", "5 - ALA mentioned", "6
- ALA published"), labels = c("data/service", "cited", "discussed", "acknowledged",
"mentioned", "published by"))


##  Main ALA data type used  - based on manual tags - by publication type

#Note when tags used: "Species occurrence records" or "Map" (in parallel to numerical tags -
details of use)
utags_articles <- str_split(dat_all_articles$Manual.Tags, "; ")
##for articles
utags_articles_det <- unlist(utags_articles)[unlist(utags_articles) %in% c("Species
occurrence records", "Map" , "Modelling", "Climate Data", "Spatial Portal", "Species lists",
"Profiles")]
count_utags_articles_det <- as.data.frame(table(utags_articles_det))
count_utags_articles_det <- rename(count_utags_articles_det, Tag = utags_articles_det)
count_utags_articles_det$Subset <- "articles"

##for grey
utags_grey <- str_split(dat_all_grey$Manual.Tags, "; ")
utags_grey_det <- unlist(utags_grey)[unlist(utags_grey) %in% c("Species occurrence records",
"Map", "Modelling", "Climate Data", "Spatial Portal", "Species lists", "Profiles")]
count_utags_grey_det <- as.data.frame(table(utags_grey_det))
count_utags_grey_det <- rename(count_utags_grey_det, Tag = utags_grey_det)
count_utags_grey_det$Subset <- "grey"

#merge into a single dataframe
count_utags_df <- rbind(count_utags_grey_det, count_utags_articles_det)
count_utags_df$Subset <- factor(count_utags_df$Subset, levels = c("grey","articles"))
#reorder levels for plotting
count_utags_df$Tag <- factor(count_utags_df$Tag, levels = c("Species lists", "Profiles",
"Spatial Portal", "Modelling", "Climate Data", "Map", "Species occurrence records"))

#plot
figure3B <- ggplot(count_utags_df, aes(x = reorder(Subset, Freq), y = Freq, fill = Tag)) +
  geom_bar(position = "fill", stat = "identity") +
  coord_flip() +
  xlab("")  +
  ylab("proportion") +
  scale_fill_manual(name = "used ALA data/service", values = Colors_7light, breaks =
c("Species lists", "Profiles", "Spatial Portal", "Modelling", "Climate Data", "Map",
"Species occurrence records"), labels = c("Species lists", "Profiles", "Spatial Portal",
"Modelling", "Climate Data", "Map", "Species occurrence records"))

## Australian focus overall - by publication type
#as.data.frame(table(dat$article_grey, dat$article_grey))
count_AU_article_grey <- count(dat_all, article_grey, AU_title_abs, name = "Count")

#plot
figure3C <- ggplot(count_AU_article_grey, aes(x = reorder(article_grey, Count), y = Count,
fill = AU_title_abs)) +
  geom_bar(position = "fill", stat = "identity") +
  coord_flip() +
  xlab("") +
  ylab("proportion") +

  scale_fill_manual(name = "Australia mentioned", values = Colors_TF)

## Community/Citizen Science focus/mentioned - by publication type
#as.data.frame(table(dat$article_grey, dat$article_grey))
count_CS_article_grey <- count(dat_all, article_grey, CS_title_abs, name = "Count")

#plot
figure3D <- ggplot(count_CS_article_grey, aes(x = reorder(article_grey, Count), y = Count,
fill = CS_title_abs)) +
  geom_bar(position = "fill", stat = "identity") +
  coord_flip() +
  xlab("") +
  ylab("proportion") +
  scale_x_discrete(labels = c("grey", "articles")) +
    scale_fill_manual(name = "community mentioned", values = Colors_TF)
```

Assemble into a multipanel figure and save as pdf.

▼ Code

```
#assemble the panels
figure3 <- figure3A / figure3B / figure3C / figure3D +
  plot_layout(ncol = 1, nrow = 4) +
  plot_annotation(tag_levels = "A")
#save the figure
ggsave(plot = figure3, here("plots", "plot.scopes_articles_grey.pdf"), width = 16, height =
16, units = "cm", dpi = 300, scale = 1.6)
```

Figure 4 - global country collaboration.

Chord diagram of top country coauthorships.
Note: this is a chord diagram is an atypical grid (base) graphics and is hard to combine in a single figure with a ggplot graphics.

▼ Code

```
load(file = here("data", "results_OpenAlex_full.RData")) #loads block_test data object with
full OpenAlex block query output
#length(block_test) #loaded data object length = 2835
#extract two variables from the results
res_OpenA <- lapply(block_test, "[", c("doi", "AU_CO"))
res_OpenA <- as.data.frame(do.call(rbind, res_OpenA))

## Processing affiliation country information
doi_country_df <- data.frame(doi = tolower(gsub("https://doi.org/", "", res_OpenA$doi)),
AU_CO = res_OpenA$AU_CO)
doi_country_all <- unlist(strsplit(doi_country_df$AU_CO, ";")) #extract country names as a
single vector
#table(doi_country_all) #table of country frequencies (note many NA)
#dim(table(doi_country_all)) #111 different country names (including NA)
doi_country_all <- doi_country_all[doi_country_all != "NA"] #remove "NA" affiliations
#length(doi_country_all) #total 12209 country values

##Count number of authors per country
country_count <- data.frame(table(doi_country_all))
#str(country_count)
country_count <- arrange(country_count, -Freq)
top10countries <- country_count[1:10,'doi_country_all']

#create big matrix of coauthorship links among all countries (incl NA for now)
doi_country_df$DB <- "SCOPUS" #add extra column to enable processing with biblioNetwork
function below
NetMatrix <- biblioNetwork(doi_country_df, analysis = "collaboration", network =
"countries", sep = ";")
#str(NetMatrix)
net_matrix <- as.matrix(NetMatrix)
#country_sum <- data.frame(countries = rownames(net_matrix), sum = rowSums(net_matrix) #sum
of links per country
small_matrix <- net_matrix[rownames(net_matrix) %in% top10countries, colnames(net_matrix)
%in% top10countries] #reduce matrix to 10 most productive countries (with highest total
number of authors)
diag(small_matrix) <- 0 #get rid of collaboration with same country

# getting rid of lower triangle (as this is duplication of info)
small_matrix[lower.tri(small_matrix)] <- 0
colnames(small_matrix) <- str_to_title(colnames(small_matrix)) #change to tile case
rownames(small_matrix) <- str_to_title(rownames(small_matrix)) #change to title case

#Fix abbreviated country names :
colnames(small_matrix)[colnames(small_matrix) == "Usa"] <- "USA"
rownames(small_matrix)[rownames(small_matrix) == "Usa"] <- "USA"
colnames(small_matrix)[colnames(small_matrix) == "United Kingdom"] <- "United Kingdom"
rownames(small_matrix)[rownames(small_matrix) == "United Kingdom"] <- "United Kingdom"

## plot country collaborations as a chord diagram for top 10 countries (using circlize
package) (colors are generated randely each time)
pdf(file = here("plots", "plot.chord_country_coauthorship.pdf"), width=6, height=6,
pointsize=14)
par(mar=c(0,0,0,0))
par(oma=c(0,0,0,0))
circos.clear()
circos.par(canvas.xlim = c(-0.8, 0.8), canvas.ylim = c(-1.0, 1.0)) #set margins
chordDiagramFromMatrix(small_matrix) #actual plotting
dev.off()
```

Figure 5 - journal impact factor, popularity, regionality.

Plot of journal Impact Factors.

▼ Code

```
dat_articles_IF <- readr::read_csv(file = here("data","dat_articles_IF.csv"))

#plot article count versus journal impact factor
figure5A <- dat_articles_IF %>%
  filter(!is.na(Journal_IF)) %>%
  count(Journal_IF) %>%
  ggplot(
      aes(x = (Journal_IF),
          y = n)) +
  geom_jitter(position = position_jitter(width = 0.3, height = 0.1), alpha = 0.5, color =
"#66c2a5") +
  xlab("Journal Impact Factor") +
  ylab("Article count")
```

Plot of most common journals.

▼ Code

```
dat_articles_IF %>%
  count(Journal_name) %>% slice_max(order_by = n, n = 15) ->  count_Journal_name #get top10
most common journal names

figure5B <- ggplot(count_Journal_name, aes(x = reorder(Journal_name, n), y = n)) +
  geom_bar(stat = "identity", fill = "#66c2a5") +
  coord_flip() +
  theme(plot.margin = margin(0, 0, 0, 0))+
  xlab("Most represented journals") +
  ylab("Article count")
```

Plot of most common journal regions/country and quartile.

▼ Code

```
dat_articles_IF_SJR <- readr::read_csv(file = here("data","dat_articles_IF_SJR.csv"))
dat_articles_IF_SJR$SJR_Best_Quartile <- dat_articles_IF_SJR$`SJR Best Quartile`
dat_articles_IF_SJR$SJR_Best_Quartile[dat_articles_IF_SJR$SJR_Best_Quartile == "-"] <- NA
dat_articles_IF_SJR$Regions <- as.factor(dat_articles_IF_SJR$Region)
levels(dat_articles_IF_SJR$Regions) <- c("Other", "Other", "Other", "Other", "Other",
"Northern America", "Pacific Region", "Western Europe")

#using package ggmosaic
figure5C <- dat_articles_IF_SJR %>% filter(!is.na(SJR_Best_Quartile)) %>% ggplot() +
  geom_mosaic(aes(x = product(Regions, SJR_Best_Quartile), fill = Regions)) +
  theme_mosaic() +
  xlab("SJR journal best Quartile") +
  ylab("journal Region") +
  theme(legend.position = "none", plot.margin = margin(0, 0, 0, 0))
```

Assemble into a multipanel figure and save as pdf.

▼ Code

```
#assemble the panels
figure5 <- figure5A / figure5B / figure5C +
  plot_layout(ncol = 1, nrow = 3) +
  plot_annotation(tag_levels = "A") +
  plot_annotation(theme = theme(plot.margin = margin()))
#save the figure
ggsave(plot = figure5, here("plots", "plot.journal_characteristics_rev.pdf"), width = 16,
height = 18, units = "cm", dpi = 300, scale = 1.6)
```

Figure 6 - open access status.

Plot article and journals Open Access status info.
Note: Meta-data available on https://unpaywall.org/data-format.

▼ Code

```
#dat_articles_OA <- readr::read_csv(file = here("data","OA_dat_articles_refs.csv")) #not
working
#dat_articles_OA <- readr::read_csv(file = here("data","dat_articles_OA_IF.csv"))
dim(dat_articles_OA)
names(dat_articles_OA)
table(is.na(dat_articles_OA$is_oa)) #no information for 595 articles

#create new data frame of counts
count_dat_articles_OA <- count(dat_articles_OA, is_oa, oa_status, name = "Count", sort =
TRUE) %>% mutate(oa_status = str_replace(oa_status, "NA", "no information"))
#count_dat_articles_OA$is_oa <- factor(as.character(count_dat_articles_OA$is_oa))
count_dat_articles_OA$is_oa[is.na(count_dat_articles_OA$is_oa)] <- "no information"
count_dat_articles_OA$oa_status[is.na(count_dat_articles_OA$oa_status)] <- "no information"
count_dat_articles_OA$is_oa <- as.factor(count_dat_articles_OA$is_oa)
levels(count_dat_articles_OA$is_oa) <- c("closed", "no information", "open access")
count_dat_articles_OA$oa_status <- factor(count_dat_articles_OA$oa_status, levels=c("no
information", "closed", "hybrid", "gold", "bronze", "green"))

# Plot as counts with Custom Color Palette (color-blind friendly) and labels
Colors_OA <- c("no information" = "#999999", "closed" = "#000000", "gold" = "#E69F00",
"green" = "#009E73", "bronze" = "#D55E00", "hybrid" = "#56B4E9")

figure6A <-
  count_dat_articles_OA %>%
  arrange(Count) %>%    # First sort by val. This sort the dataframe but NOT the factor
levels
  mutate(is_oa2 = factor(is_oa, levels = c("no information", "closed", "open access"))) %>%
# This trick update the factor levels
  ggplot(aes(x = is_oa2, y = Count, fill = oa_status)) +
  geom_bar(stat = "identity") +
  coord_flip() +
  xlab("") +
  scale_fill_manual(values = Colors_OA, name = 'open access type:', breaks = c('green',
'bronze', 'gold', 'hybrid'), labels = c("Green = unpublished version freely available via a
repository", "Bronze = published version is free to read only on the publisher's website",
"Gold = author pays for a free-to-read version to be available from the fully-OA journal",
"Hybrid = author pays for a free-to-read version to be available from the subscription
journal")) +
  geom_text(aes(label = Count), color="white", size = 3, position = position_stack(vjust =
0.5)) +
    theme(legend.position = "bottom", legend.direction = "vertical") +
  ylab("article count")

#Note: this plot matches this table:
# table(dat_articles_OA$oa_status, dat_articles_OA$is_oa, useNA = "always") #OA status of
articles
# table(dat_articles_OA$is_oa)/sum(table(dat_articles_OA$is_oa)) #percentages for OA TRUE or
FALSE
```

Plot of OA status - counts across publication years.

▼ Code

```
dat_articles_OA %>%
  count(Publication.Year, is_oa) %>% drop_na() -> count_year_oa

#assign custom colors
Colors_OA2 <- c("#f5d4fc", "#000000")

#plot change of proportion oa across years
figure6B <- ggplot(count_year_oa, aes(x = as.character(Publication.Year), y = n, fill =
reorder(is_oa, n))) +
  geom_bar(stat = "identity", position = "fill") + #, position = "dodge"
  xlab("") +
  scale_fill_manual(values = rev(Colors_OA2), name = "", labels = c("closed access", "open
access")) +
  ylab("proportion") +
  legend_bottom()
```

Assemble into a multipanel figure and save as pdf.

▼ Code

```
#assemble the panels
figure6 <- figure6A / figure6B +
  plot_layout(ncol = 1, nrow = 2) +
  plot_annotation(tag_levels = "A")
#save the figure
ggsave(plot = figure6, here("plots", "plot.oa_status.pdf"), width = 16, height = 12, units =
"cm", dpi = 300, scale = 1.6)
```

New Figure 2 (added after manuscript revision).
Note this figure is generated using internal files from The ALA and data gathered from GBIF via API.

▼ Code

```
setwd("C:/Users/WAI045/OneDrive - CSIRO/ALA/Citations Tracker")

##### Get Datasets #####
dataset_df <- request("https://api.gbif.org/v1/dataset/search/export?
publishingCountry=AU&format=CSV") |>
  req_perform() |>
  resp_body_string() |>
  read_csv()

##### Takes ~2hrs to run #####
# all_papers <- map(
#   .x = 1:nrow(dataset_df),
#   .f = function(x) {
#     lit_counts <- lit_count(datasetKey = dataset_df$dataset_key[x])
#     offset_seq <- seq(0, lit_counts, by = 500)
#
#     this_dataset <- map(
#       .x = 1:length(offset_seq),
#       .f = function(y) {
#         lit_api <- paste0("https://api.gbif.org/v1/literature/search?gbifDatasetKey=",
#                           dataset_df$dataset_key[x],
#                           "&limit=500&offset=",
#                           offset_seq[y])
#         lit_df <- (request(lit_api) |>
#               req_perform() |>
#               resp_body_json())$results
#         Sys.sleep(2)
#         return(lit_df)
#       }) |>
#       list_flatten() |>
#       map(.f = function(y) {
#             paper_df <- y |> tibble() |> t() |> data.frame() |> tibble()
#             colnames(paper_df) <- names(y)
#             return(paper_df)}) |>
#       list_rbind() |>
#       mutate(dataset = dataset_df$title[x])
#   },
#   .progress = TRUE
# ) |>
#   list_rbind()
#
# save(all_papers, file = "data/all_papers.rds")
load("data/all_papers.rds")

##### Clean Dataframe #####
all_papers_clean <- all_papers |>
  select(-gbifDownloadKey, -gbifOccurrenceKey, -gbifNetworkKey, -gbifProjectIdentifier, -
gbifProgramme, -tags) |>
  rowwise() |>
  mutate(
    across(-c(authors, dataset),
           function(x) unlist(x) |> paste0(collapse = " | "))) |>
  mutate(authors = map(.x = authors,
                       .f = function(x) paste0(x$lastName, ", ", x$firstName)) |>
           unlist() |> paste0(collapse = "; ")) |>
  distinct() |>
  mutate(true_holder = TRUE) |>
  pivot_wider(names_from = dataset, values_from = true_holder, values_fill = FALSE) |>
  mutate(across(everything(),
                function(x) {ifelse(x %in% c("-", ""), NA, x)})) |>
  filter(publisher != "IUCN" | literatureType != "REPORT")

write_csv(all_papers_clean, file = "data/all_papers_clean.csv")

##### Compare to ALA citations #####
ALA_zotero <- read_csv("data/ala_zotero_data.csv") |>
  mutate(across(everything(),
                function(x) {ifelse(x %in% c("-", ""), NA, x)}))

joined_papers <- all_papers_clean |>
  select(year, identifiers, literatureType) |>
  rename(`Publication Year` = year,
         DOI = identifiers,
         `Item Type` = literatureType) |>
  mutate(`Citation Source` = "GBIF") |>
  rbind(
    ALA_zotero |>
      select(`Publication Year`, `DOI`, `Item Type`) |>
      mutate(`Citation Source` = "ALA")
  ) |>
  mutate(`Item Type` = to_title_case(`Item Type`),
         `Item Type` = case_when(
           `Item Type` == "Journal" ~ "Journal Article",
           `Item Type` == "Webpage" ~ "Web Page",
           .default = `Item Type`
         )) |>
  distinct()

joined_papers_clean <- joined_papers |>
  filter(!is.na(DOI)) |>
  mutate(true_holder = TRUE) |>
  pivot_wider(names_from = `Citation Source`, values_from = true_holder, values_fill =
FALSE) |>
  mutate(`Citation Source` = case_when(
    GBIF & !ALA ~ "GBIF",
    !GBIF & ALA ~ "ALA",
    GBIF & ALA ~ "Both",
    .default = NA)) |>
  mutate(`Publication Year` = as.numeric(`Publication Year`),
         `Citation Source` = factor(`Citation Source`, levels = c("GBIF", "Both", "ALA")))

citation_plot <- ggplot(joined_papers_clean) +
  geom_bar(aes(x = `Publication Year`, fill = `Citation Source`)) +
  scale_fill_manual(values = c("ALA" = "#FFA300", "Both" = "#B43CC8", "GBIF" = "#509E2F")) +
  theme_classic() +
  ylab("Number of Citations")

ggsave(filename = "figures/citation_plot.pdf")

#### DOIs with different years
# 10.1016/j.biocon.2021.109425
# 10.1016/j.cub.2020.10.053
# 10.1016/j.ecochg.2022.100062
# 10.1088/1748-9326/aaf5db

##### Total GBIF Papers (- ALA ones) #####
GBIF_AU_totals <- all_papers_clean |>
  filter(!is.na(identifiers)) |>
  select(year, identifiers, literatureType) |>
  rename(`Publication Year` = year,
         DOI = identifiers,
         `Item Type` = literatureType) |>
  mutate(`Publication Year` = as.numeric(`Publication Year`)) |>
  count(`Publication Year`)

##### Takes ~40min to run #####
# GBIF_all_papers <- map(
#   .x = 2007:2024,
#   .f = function(x) {
#     lit_counts <- lit_count(datasetKey = dataset_df$dataset_key[x])
#     offset_seq <- seq(0, lit_counts, by = 800)
#
#     this_dataset <- map(
#       .x = 1:length(offset_seq),
#       .f = function(y) {
#         lit_api <- paste0("https://api.gbif.org/v1/literature/search?year=",
#                           x,
#                           "&limit=800&offset=",
#                           offset_seq[y])
#         lit_df <- (request(lit_api) |>
#                      req_perform() |>
#                    resp_body_json())$results
#         Sys.sleep(2)
#         return(lit_df)
#       },
#       .progress = TRUE
#     ) |>
#       list_flatten() |>
#       map(.f = function(y) {
#         paper_df <- y |> tibble() |> t() |> data.frame() |> tibble()
#         colnames(paper_df) <- names(y)
#         return(paper_df)}) |>
#       list_rbind()
#   }
# ) |>
#   list_rbind()
#
# save(GBIF_all_papers, file = "data/GBIF_all_papers.rds")
load("data/GBIF_all_papers.rds")

GBIF_all_papers_clean <- GBIF_all_papers |>
  select(-gbifDownloadKey, -gbifOccurrenceKey, -gbifNetworkKey, -gbifProjectIdentifier, -
gbifProgramme, -tags) |>
  rowwise() |>
  mutate(
    across(-c(authors),
           function(x) unlist(x) |> paste0(collapse = " | "))) |>
  mutate(authors = map(.x = authors,
                       .f = function(x) paste0(x$lastName, ", ", x$firstName)) |>
           unlist() |> paste0(collapse = "; ")) |>
  distinct() |>
  mutate(across(everything(),
                function(x) {ifelse(x %in% c("-", ""), NA, x)})) |>
  filter(publisher != "IUCN" | literatureType != "REPORT") |>
  filter(!(id %in% all_papers_clean$id))

write_csv(all_papers_clean, file = "data/GBIF_all_papers_clean.csv")

GBIF_totals <- GBIF_all_papers_clean |>
  filter(!is.na(identifiers)) |>
  select(year, identifiers, literatureType) |>
  rename(`Publication Year` = year,
         DOI = identifiers,
         `Item Type` = literatureType) |>
  count(`Publication Year`)

GBIF_citation_plot <- rbind(
  GBIF_totals |> mutate(ALA_data = "don't use ALA data"),
  GBIF_AU_totals |> mutate(ALA_data = "use ALA data")) |>
  ggplot() +
  geom_bar(aes(x = `Publication Year`, y = n, fill = ALA_data), stat = "identity") +
  scale_fill_manual(name = "GBIF citations that...",
                    values = c("use ALA data" = "#FFA300", "don't use ALA data" =
"#509E2F")) +
  theme_classic() +
  ylab("Number of Citations")

ggsave(filename = "figures/GBIF_citation_plot.pdf")


##### Stage 2 #####

rbind(
  all_papers_clean |>
    select(authors, added, published, identifiers, keywords, literatureType, openAccess,
peerReview, source, title, topics, year, abstract, citationType) |>
    mutate(category = "GBIF: cites ALA"),
  GBIF_all_papers_clean |>
    select(authors, added, published, identifiers, keywords, literatureType, openAccess,
peerReview, source, title, topics, year, abstract, citationType) |>
    mutate(category = "GBIF: doesn't cite ALA")) |>
  # mutate(topics = str_split(topics, " \\| "),
  #        keywords = str_split(keywords, " \\| ")) |>
  write_csv("data/GBIF_citations_all.csv")


##### Plotting New Figures #####
all_citations_totals <- joined_papers_clean |>
  mutate(
    `Citation Source` = as.character(`Citation Source`),
    `Citation Source` = ifelse(`Citation Source` == "Both", "ALA", `Citation Source`),
    `Citation Source` = ifelse(`Citation Source` == "ALA", "AU data via ALA", "AU data via
GBIF")) |>
  count(`Publication Year`, `Citation Source`) |>
  rbind(
    GBIF_totals |>
      mutate(`Citation Source` = "GBIF without AU data") |>
      select(`Publication Year`, `Citation Source`, n)
  ) |>
  mutate(
    `Publication Year` = as.numeric(`Publication Year`),
    `Citation Source` = factor(`Citation Source`, levels = c("GBIF without AU data", "AU
data via GBIF", "AU data via ALA"))) |>
  filter(`Publication Year` <= 2022)

all_citations_plot <- ggplot(all_citations_totals) +
  geom_bar(aes(x = `Publication Year`, y = n, fill = `Citation Source`),
           position = "stack", stat = "identity") +
  scale_fill_manual(values = c("GBIF without AU data" = "#8fcceb",
                               "AU data via GBIF" = "#239e60",
                               "AU data via ALA" = "#83d166")) +
  theme_bw() +
  ylab("Number of Citations")
ggsave("figures/citation_plot_cw.pdf", all_citations_plot, width = 8, height = 6, units =
"in")
pdf_convert("figures/citation_plot_cw.pdf",
            format = "png", dpi = 500,
            filenames = "figures/citation_plot_cw.png")

##### Tree Map #####
treemap_data <- ALA_zotero |>
```

```
rename(item_type = "Item Type") |>
mutate(
  item_type = case_when(
    item_type == "manuscript" ~ "journalArticle",
    item_type %in% c("book", "bookSection") ~ "bookOrBookSection",
    item_type %in% c("magazineArticle", "encyclopediaArticle") ~ "magazineOrEncyclopedia",
    item_type %in% c("artwork", "videoRecording") ~ "mixedMedia",
    .default = item_type
  )
) |>
count(item_type) |>
mutate(
  group_type = case_when(
    item_type %in% c("blogPost", "document", "magazineOrEncyclopedia",
                     "mixedMedia", "newspaperArticle", "report", "webpage") ~ 2,
    .default = 1
  ) |>
    factor(levels = c(1, 2),
           labels = c("Scientific content",
                      "Grey lit & cultural content")),
  item_type = to_sentence_case(item_type)) |>
mutate(item_type = gsub("\\s", "\n", item_type)) |>
mutate(
  item_type = case_when(
    item_type == "Book\nor\nbook\nsection" ~ "Book\nor\nBook section",
    item_type == "Magazine\nor\nencyclopedia" ~ "Magazine\nor\nEncylopedia",
    .default = item_type
  )) |>
arrange(group_type, desc(n)) |>
mutate(fill_col = c("#d8f5ce",
                    "#b0e69c",
                    "#98e37d",
                    "#83d166",
                    "#6bba4e",
                    "#5ba63f",
                    "#f5e9ce",
                    "#e7d09d",
                    "#e3c57d",
                    "#d1b166",
                    "#ba9a4f",
                    "#a6873f",
                    "#8a6c28"))

types_treemap <- ggplot(treemap_data,
                        aes(area = n,
                            fill = fill_col,
                            subgroup = group_type,
                            label = item_type)) +
  geom_treemap() +
  scale_fill_identity() +
  geom_treemap_subgroup_border() +
  geom_treemap_text(
    padding.x = grid::unit(2, "mm"),
    padding.y = grid::unit(3, "mm")) +
  theme(legend.position = NULL)

ggsave("figures/paper_type_treemap.pdf", types_treemap, width = 5, height = 5, units = "in")
pdf_convert("figures/paper_type_treemap.pdf",
            format = "png", dpi = 500,
            filenames = "figures/paper_type_treemap.png")
```