

# Evaluating institutional open access performance: Methodology, challenges and assessment

August 12, 2020

Chun-Kai Huang<sup>1</sup>, Cameron Neylon<sup>1,2</sup>, Richard Hosking<sup>2</sup>, Lucy Montgomery<sup>1,2</sup>, Katie Wilson<sup>1</sup>, Alkim Ozaygen<sup>1</sup>, Chloe Brookes-Kenworthy<sup>1</sup>

<sup>1</sup>*Centre for Culture and Technology, School of Media, Creative Arts and Social Inquiry, Curtin University, Kent St, Bentley 6102, Western Australia* <sup>2</sup>*Curtin Institute for Computation, Kent St, Bentley 6102, Western Australia*

## Abstract

Open access to research outputs is becoming rapidly more important to the global research community and society. Changes are driven by funder mandates, institutional policy, grass-roots advocacy and culture change. It has been challenging to provide a robust, transparent and updateable analysis of progress towards open access that can inform these interventions, particularly at the institutional level. Here we propose a set of essential requirements for such a process and present a large-scale analysis of open access progress across 1,207 institutions world-wide that shows substantial progress being made. The analysis detects responses that coincide with policy and funding interventions. Among the striking results are the high performance of Latin American and African universities, particularly for gold open access, whereas overall open access levels in Europe and North America are driven by repository-mediated access. Our analysis shows the world's leading institutions achieving around 80% open access for publications in 2017.

Keywords: Open Access; Evaluation Framework; Unpaywall; Microsoft Academic; Web of Science; Scopus.

## 1 Introduction

Open access is a policy aspiration for research funders, organisations, and communities globally. While there is substantial disagreement on the best route to achieve open access, the idea that wider availability of research outputs should be a goal is broadly shared. Over the past decade, there has been massive increase in the volume of publications available open access. Piwowar et al. (2018) showed that the global proportion of open access articles was about 45% for those published in 2015, compared to around 5% before 1990. A more recent projection suggests that 44% of all outputs ever published will be freely accessible in 2025 (Piwowar et al., 2019).

This massive increase has been driven in large part by policy initiatives. Medical research funders such as the Wellcome Trust and Medical Research Council in the UK and the National Institutes of Health in the US led a wide range of funder policy interventions. Universities such as Harvard, Liege, Southampton and others developed local policies and infrastructures that became more widely adopted. Plan S, led by a coalition of funders, announced in 2019 has as its goal the complete

conversion of scholarly publishing to immediate open access. This is the most ambitious, and therefore the most controversial, policy initiative to date with questions raised about the approach (Rabesandratana, 2019; Haug, 2019; Barbour and Nicholls, 2019), implementation details (McNutt, 2019; Gómez-Fernández, 2019; Brainard, 2019; Agustini and Berk, 2019), and unintended side effects for existing programs outside North America and Northwestern Europe (Debat and Babini, 2019; Aguado-López and Becerril-García, 2019).

A recent report by Larivière and Sugimoto (2018) showed a link between the monitoring of policy and its effectiveness, describing strong performance by articles supported by funders that had implemented monitoring and compliance checks for their policies. By comparison open access for works funded by Canadian funders, which did not monitor compliance, were shown to lag substantially even when disciplinary effects were taken into account.

There is a need for critical and inclusive evaluation of open access performance that can address regional and political differences. For example, the SciELO project has successfully implemented an electronic publishing model for journals resulting in a surge of journal-mediated open access (Packer, 2009; Wang et al., 2018). Recent work by Iyandemeye and Thomas (2019) showed that, for biomedical research, there was a greater level of open access for articles published from countries with a lower GDP, particularly for those in sub-Saharan Africa. This provides evidence of national or regional effects on publication cultures that lead to open access. Meanwhile, Siler et al. (2018) showed that, for the field of Global Health, lower-ranked institutions are more likely to publish in closed outlets. They suggest this is due to the cost of article processing charges showing the importance of considering institutional context when examining open access performance.

Despite the scale and success (at least in some areas) of policy interventions, there is limited comparative and quantitative research about which policy interventions have been the most successful. In part this is due to a historical lack of high-quality data on open access, the heterogeneous nature of the global scholarly publishing endeavour, and the consequent lack of any baseline against which to make comparisons.

## 1.1 Aim of Study

We have argued (Montgomery et al., 2018) that the key to understanding and guiding the cultural changes that underpin a transition to openness is analysis at the level of research institutions. While funders, national governments, and research communities create the environments in which researchers operate, it is within their professional spaces that choices around communication, and their links to career progression and job security are strongest. Analysis of how external policy leads to change at the level of universities is critical. However, providing accurate and reliable data on open access at the university level is a challenge.

The most comprehensive work on open access at the university level currently available is that included in the CWTS Leiden Ranking (Robinson-Garcia et al., 2019; 2020). This utilises an internal Web of Science database and data from Unpaywall to provide estimates of open access over a range of timeframes. These data have highlighted the broad effects of funder policies (notably the performance of UK universities in response to national policies) while also providing standout examples from regions that are less expected (for instance Bilkent University in Turkey).

A concern in any university evaluation is the existing disciplinary bias in large bibliographic sources used to support rankings. For example, the coverages of Web of Science and Scopus were shown to be biased toward the sciences and the English language. We, and others have shown how sources

can be biased towards disciplines and languages (Mongeon and Paul-Hus, 2016) and how evaluation frameworks based on single sources of output data can provide misleading results (Huang et al., 2020a). In a companion white paper to this article we provide more details of these issues with a sensitivity analysis of the data presented here (Huang et al., 2020b). If we are to make valid comparisons of universities across countries, regions and funders to examine the effectiveness of open access policy implementation there is a critical need for evaluation frameworks that provide fair, inclusive, and relevant measurement of open access performance.

Alongside coverage of data sources are issues of scope (which institutions, what set of objects), metrics (numbers or proportions) and data completeness. Our pragmatic assessment is that any evaluation framework should be tied to explicit policy goals and be shaped to deliver that. Following from our work on open knowledge institutions (Montgomery et al., 2018) our goals in conducting an evaluation exercise and developing the framework are as follows:

1. Maximising the amount of research content that is accessible to the widest range of users, in the first instance focusing on existing formal research content for which metadata quality is sufficiently high to enable analysis
2. Developing an evaluation framework that drives an elevation of open access and open science issues to a strategic issue for all research-intensive universities
3. Developing a framework that is sensitive to and can support universities taking a diversity of approaches and routes towards delivering on those goals

In terms of a pragmatic approach to delivering on these we therefore intend to:

1. Focus on research-intensive institutions, using existing rankings as a sample set
2. Seek to maximise the set of objects which we can collect and track while connecting them to institutions (i.e., increase recall but not at the expense of precision)
3. Focus on proportions of open access as a performance indicator rather than absolute numbers
4. Publicly report on the details of performance for high performing institutions (and provide strategic data on request to others)
5. Report on the diversity of paths being taken to deliver overall access by a diverse group of universities
6. Develop methodology that is capable of identifying which policy interventions have made a difference to outcome measures and any ‘signature’ of those effects

With the above in mind, this study proposes a set of essential requirements for robust institutional open access evaluation, and presents a large-scale analysis of universities by drawing and integrating data from multiple data sources. This work differs from the CWTS Leiden Rankings by extending the coverage of research outputs beyond the Web of Science. The data workflow we have developed is also transparent, reproducible, and updatable, which makes robust and longitudinal analysis more easily attainable. We emphasise that a simple numerical ranking of universities cannot be justified given there is minimal significant difference across them. Instead, we highlight how the resulting comprehensive overview of the open access landscape and the underlying trends over time can provide deep insights on effects of policy interventions.

## 2 Limitations and further work

Our analysis process includes automated approaches for collecting the outputs related to specific universities, and the analysis of those outputs. Currently the addition of new universities, and the updating of large data sources is partly manual (see Supplementary file 1) but we also expect to

automate this in the near future. Along with the nature of this article this will provide an updatable report and longitudinal dataset that can provide a consistent and growing evidence source for open access policy and implementation analysis.

While it is clear (Huang et al., 2020b) that our analysis has limitations in its capacity to provide comparable estimates of open access status across all universities, our approach does provide a reproducible and transparent view of overall global performance. There are challenges to be addressed with respect to small universities and research organisations and we have taken a necessarily subjective view of which institutions to include (see Supplementary file 1). Our approach systematically leaves out universities with very small numbers of outputs (i.e., less than 100 outputs), and universities with very extreme open access proportions as these are the universities for which we have less statistical confidence in the results. This is also in-line with our intended focus on research-intensive universities. These small institutions are of significant interest but will require a different analysis approach. We include a country-level summary result for the full set of institutions (including small institutions left out from the main article) in our data set in Supplementary file 3. We recognise that the inclusion of universities beyond our initial selection can also potentially change the results.

We have used multiple sources of bibliographic information with the goal of gaining a more inclusive view of research outputs. Despite this there are still limitations in the coverage of these data sources, and a likely bias towards STEM disciplines. In addition the focus of Unpaywall on analysis of outputs with Crossref DOIs means that we are missing outputs for disciplines (humanities) and output types (books) where the use of DOIs is lower. While we have performed additional sensitivity analysis on the use of different data sources (Huang et al, 2020b), further work is required to understand the relationships between disciplinary biases in data sources and the influence of external policies (such as those of the National Institutes of Health and the Wellcome Trust) on repositories. In addition, due to the nature of this work and to limitations on the use of Web of Science and Scopus APIs, we have collected data from these two sources over a period of time. Although we expect such changes to be small those effects are not clearly represented in our data. For other data sources we are able to precisely define the data dump used for our analysis, supporting reproducibility as well as modified analyses. We also recognise that our longitudinal analysis is done post hoc from a fixed time point (i.e., looking back). This implies our results for a particular year may not be exactly the same as the results observed *in* that year, which are likely to be influenced by embargoes and moving paywalls.

In this article, we have drawn comparisons between publisher-mediated open access and repository-mediated open access to identify different institutional paths to open access. The overall trends are retained if we separate out outputs that are only available through repositories (i.e., not free to access via publishers) and compare these to publisher-mediated open access instead. However, we recognised that at the individual university level, this can be dependent on institutional policy and whether the corresponding data can be properly captured by current systems that determines open access status.

There is significant opportunity for improving the data sources on sets of outputs and how they can be grouped (e.g. by people, discipline, organisation, country, etc). Improvements to institutional identifier systems such as the Research Organisation Registry, increased completeness of metadata records, particularly that provided by publishers via Crossref on affiliation, ORCIDs and funders, and enhancing the coverage of open access status data (for instance by incorporating data from CORE and BASE), will all enhance coverage. There are also opportunities to expand the coverage by incorporating a wider range of bibliographic data sources.

We have sought to make our methodology and approach as reproducible and reusable as is practicable. The main challenges lie in the level of accessibility to the closed data sources (i.e., Web of Science and Scopus) and more generally with data sources which are not available in the form of identifiable snapshots (see Supplementary Methodology). In addition, semi-automated processing and manual searches are required to link institutional identifiers across multiple data sources (as discussed in detail in Supplementary file 1). For the rest of the framework, we have provided transparent descriptions and source code that allow reprocessing, as described in the next section. Further work is required to provide a complete and deployable framework for replication but we are making progress in this area (see <https://github.com/The-Academic-Observatory/observatory-platform> for an example).

### 3 Methods

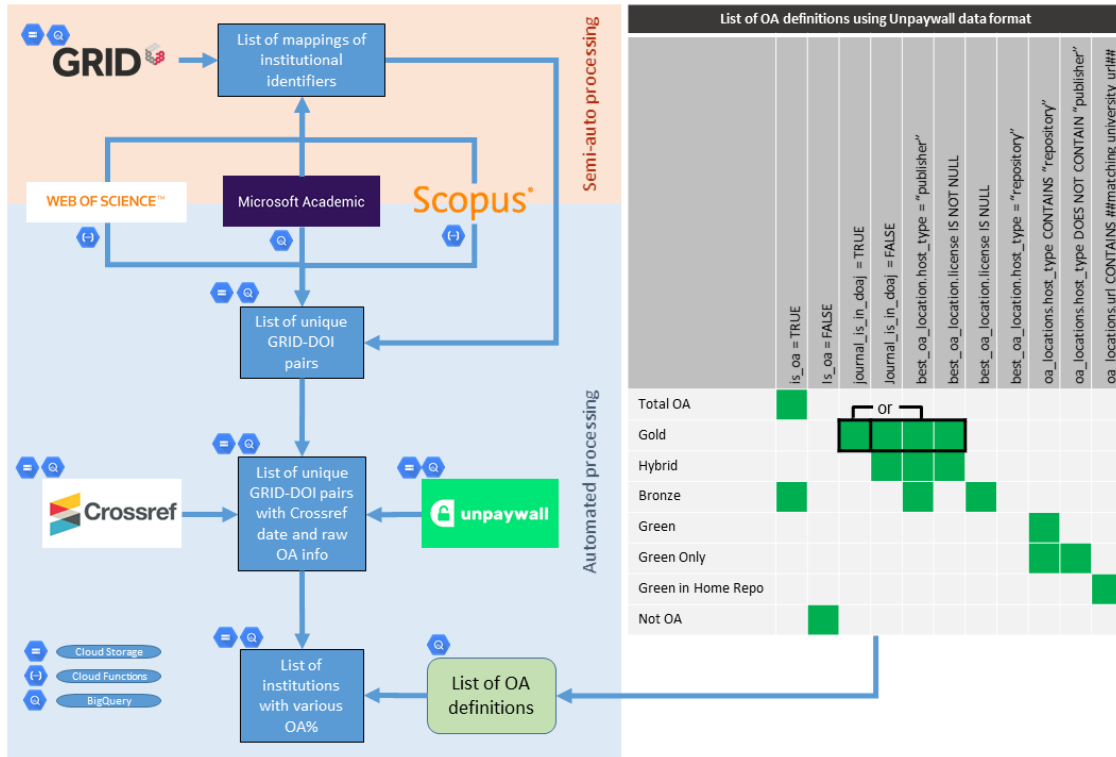
There have been many differing assessments of open access performance over the past 10-15 years. Many of the differences between these have been driven by details in the approach. This combined with limited attention to reproducibility has led to confusion and a lack of clarity on the rate and degree of progress to open access (Green, 2019). Hence, we developed a reproducible workflow capable of quantifying a wide range of open access characteristics at the institutional level. The overall workflow is summarised diagrammatically in Figure 1. This includes a mapping of open access definitions and the Unpaywall information we used to construct them. Briefly, we gather output metadata from searches in Microsoft Academic (Sinha et al., 2015; Wang et al., 2019), Web of Science and Scopus, for each university. From this full set we gather the corresponding Crossref DOIs from the metadata of each output focusing on this set. Unpaywall is consulted to determine open access status. Detailed discussions of the data sources, precise data snapshots used, and technical details of the data infrastructure can be found in Supplementary file 1. A summary of selected open access definitions is given in the table below:

OA	
Type	Description
Total OA	A research output that is free to read online, either via the publisher website or in an OA repository.
Gold	A research output that is either published in a journal listed by the Directory of Open Access Journals (DOAJ), or (if journal not in DOAJ) is free to read via publisher with any license.
Hybrid	A research output that is published in a journal not listed by DOAJ, but is free to read from publisher with any license.
Bronze	A research output that is free to read online via publisher without a license.
Green	A research output that is free to read online via an OA repository.
Green Only	A research output that is free to read online via an OA repository, but is not available for free via the publisher.
Green in Home Repo	A research output that is free to read online via the matched affiliation's institutional repository.

Further discussions of the categories of open access definitions are provided in Supplementary file 1.

We have selected to focus mainly on total open access, publisher-mediated open access (i.e., gold), and repository-mediated open access (i.e., green) due to the ease of comprehension, data quality, and ability to show potential effects of selected policies. Levels of hybrid open access and green in home repository for selected universities (ones for which we have more confidence in data) are also included to further support the analysis of policy effects.

**Figure 1: Workflow of data collection and mapping of open access definitions to Unpaywall metadata.**



As we have noted previously (Huang et al., 2020a) there is a sensitivity associated to the choices in bibliographic data sources when they are used to create a ranking. For this analysis we therefore chose to combine all three datasets (i.e., Microsoft Academic, Web of Science and Scopus). In the companion white paper (Huang et al., 2020b) we provide a comprehensive sensitivity analysis on the use of these different datasets, the use of different versions of Unpaywall, and the relations between confidence levels and sample size.

Briefly, it is our view that to provide a robust assessment of open access performance the following set of essential requirements must be met:

1. The set of outputs included in each category and a traceable description of how they were collected must be transparently described. Provided here by a description of the data sources and the procedures used to collect DOIs for each institution (see Supplementary file 1). In this article, the categories of outputs are defined by institutions but they may as well be categorised by individuals, disciplines, or countries etc.

2. A clearly defined, open and auditable data source on open access status. Provided here by a defined and identified Unpaywall snapshot (see Supplementary file 1).
3. A clearly defined and implementable description of how open access status data is interpreted. Provided here in Figure 1 and in Supplementary file 1 in the form of the SQL query used to establish open access status categories for each DOI.
4. Provision of derived data and analysis in auditable form. Provided here the derived data as open data (Huang et al., 2020c), code for the analysis of derived data as Jupyter notebooks (Huang et al., 2020d), and upstream data analysis in the form of SQL queries used (Huang et al., 2020c).

We have limited our data sharing in two ways. Firstly, we do not provide the full list of DOIs obtained from each source, due to Terms of Service restrictions. Secondly, we have not identified institutions individually except for those that fall within the top 100 globally for total open access, journal-mediated open access, or repository-mediated open access. Both derived datasets of the de-anonymised top 100 and the full dataset containing all institutions in anonymised form are made available (Huang et al., 2020c).

## 4 Results

### 4.1 Global universities in terms of total open access, gold open access and green open access

In Figure 2, we present results of universities for each of the categories of total open access, publisher-mediated open access (“gold”) and repository-mediated open access (“green”) for publications assigned to the year 2017 (see Figure 2 - figure supplement 1 for equivalent plots for 2016 and 2018). The universities are grouped by region. We have chosen to focus on 2017 here as this is the most recent year for which we have the most confidence on the completeness of data, taking into consideration of the data collection process and issues surrounding embargoes. However, we do observe consistent general patterns across results for all three years. The top 100 institutions in each of the open access categories (for 2017) are also provided in Supplementary file 2. This is, to our knowledge, the first set of university rankings that provides a confidence interval on the quantitative variable being ranked and compensates for the multiple comparisons effect (see Supplementary file 1 for details). Across this top 100 the statistical difference between universities at the 95% confidence shows that a simple numerical ranking cannot be justified.

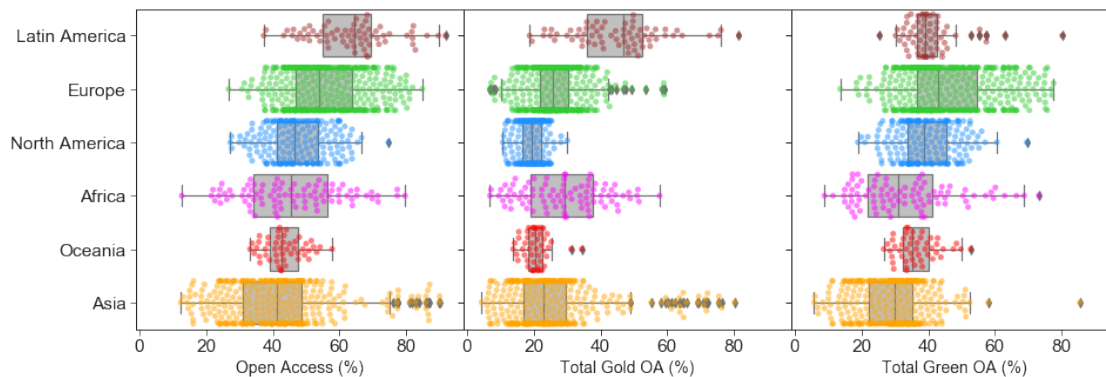
The high performance of a number of Latin American and African universities, together with a number of Indonesian universities, particularly with respect to gold open access, is striking. For Latin America this is sensitive to our use of Microsoft Academic as a data source (Huang et al., 2020b) showing the importance of an inclusive approach. The outcomes for Indonesian universities are also consistent with the latest report on country-level analysis (Van Noorden, 2019). These suggest that the narrative of Europe and the USA driving a publishing-dominated approach to open access misses a substantial part of the full global picture.

The highest performers in terms of open access via repositories are dominated by UK universities. This is not surprising given the power of the open access mandate associated with the Research Excellence Framework to drive university behaviour. It is perhaps interesting that few US universities appear in this group (with CalTech and MIT the exceptions). This suggests that while the National Institutes of Health mandate has been very effective at driving open access to the biomedical literature limited inroads have been made into other disciplines in the US context, despite the

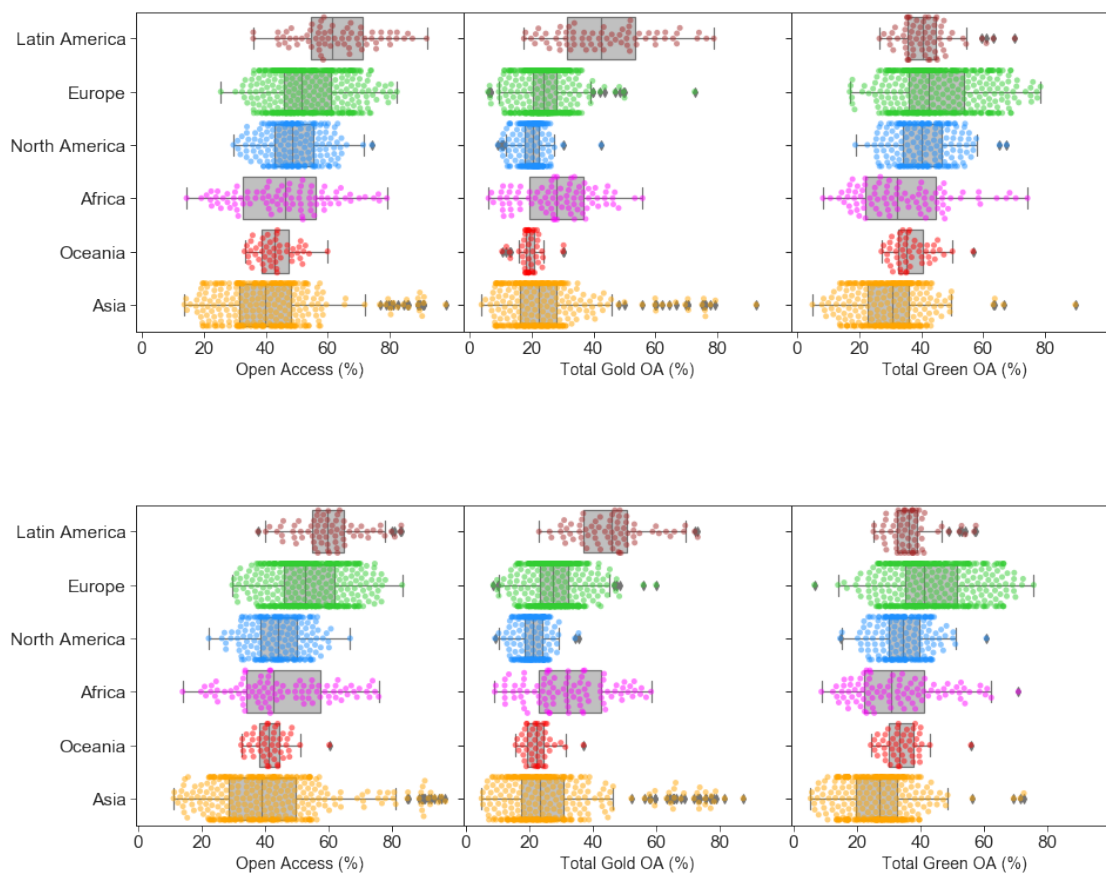
White House memorandum. As was seen in the Leiden Ranking, Bilkent University from Turkey also emerges as a stand-out performer.



**Figure 2: Percentages of institutional Total OA, Gold OA and Green OA (left to right) grouped by regions for 2017.**



**Figure 2 - figure supplement 1**



## 4.2 The global picture and its evolution

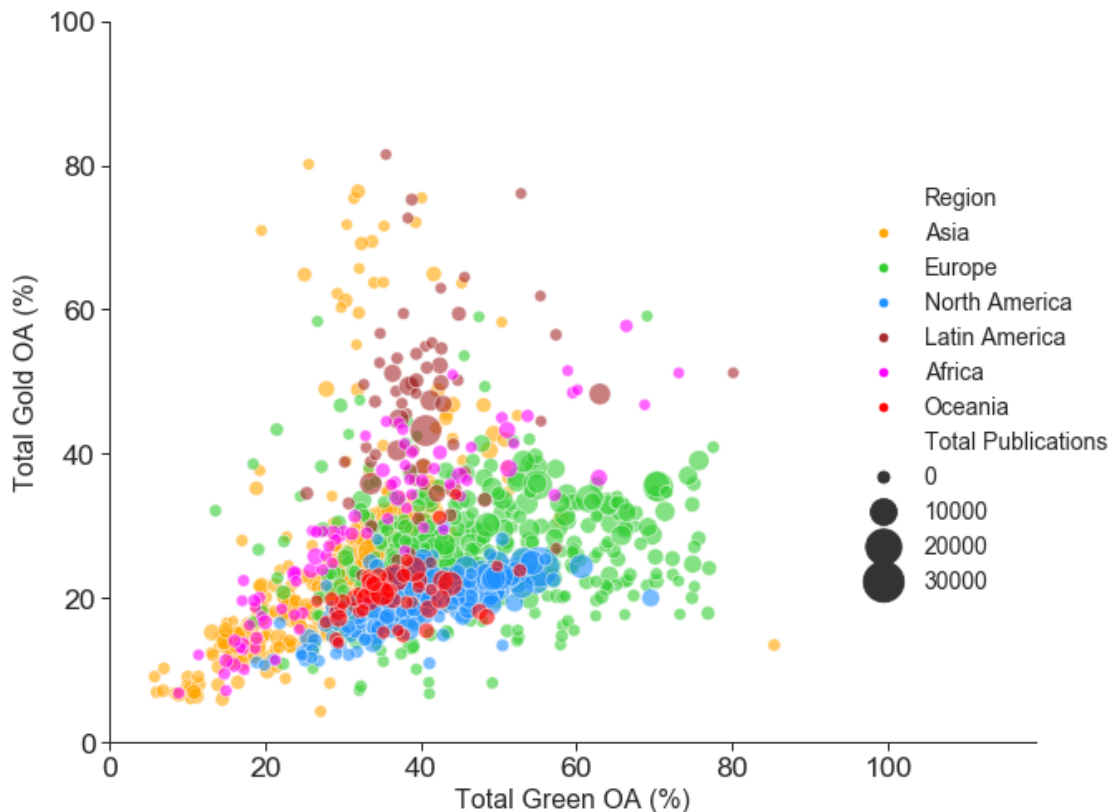
The levels of total open access, publisher-mediated open access and repository-mediated open access for 1,207 universities for publications in 2017, grouped by country, can be found in Supplementary file 3. Countries are ordered by the median total open access percentage. Amongst countries with a large number of universities in the dataset the UK is a clear leader with Indonesia, Brazil, Columbia, the Netherlands, and Switzerland showing a strong performance.

There are high performing universities in Latin America (i.e., Peru, Costa Rica, Columbia, Chile, Brazil) and Uganda as well as a range of European countries. Latin American countries owe their performance in large part to open access journals (“gold”) whereas European countries see a more significant contribution from repository based open access (“green”). Many countries have universities that are high performers in terms of the proportion of open access, while overall country performance can be linked to policy mandates and infrastructure provision.

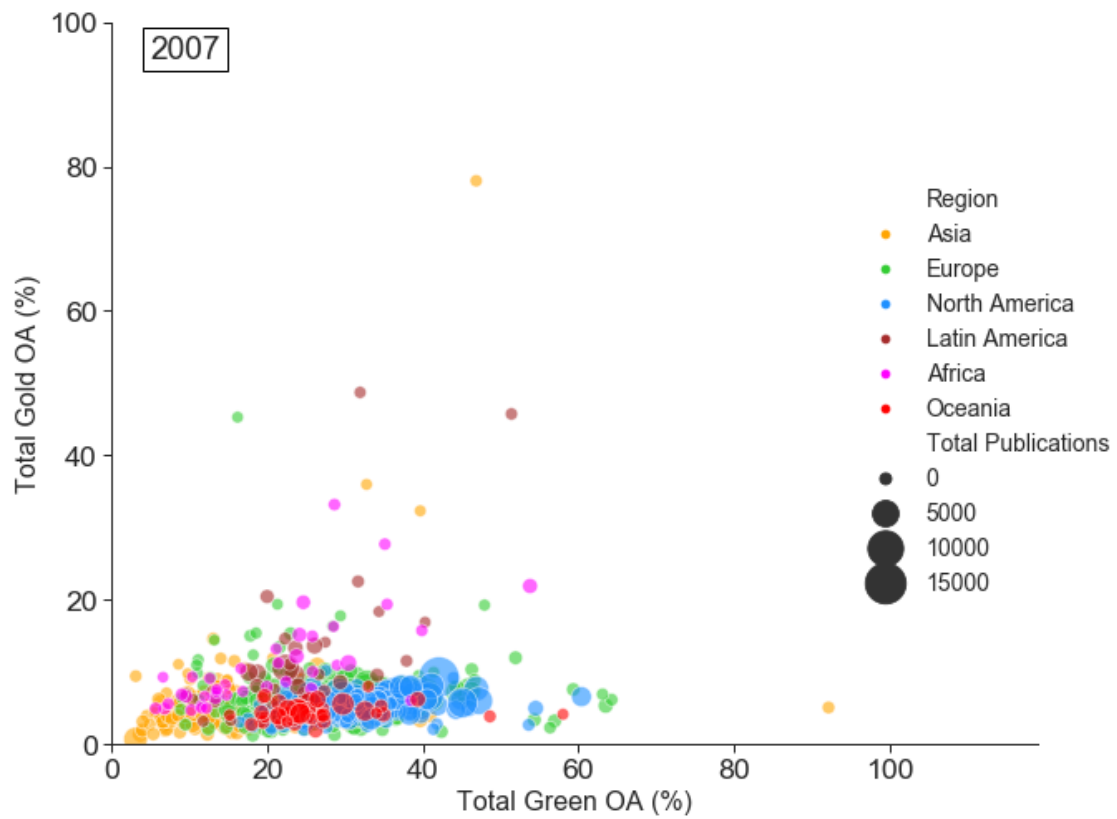
To examine the global picture for the 1,207 universities in our dataset and to interrogate different paths to open access we plot the overall level of repository mediated (“green”) and publisher mediated (“gold”) open access for each university over time coloured by region as previously. Figure 3 presents the results for 2017 (with changes over time shown in the animated version, see also Figure 3 - figure supplement 1 output for each year individually).

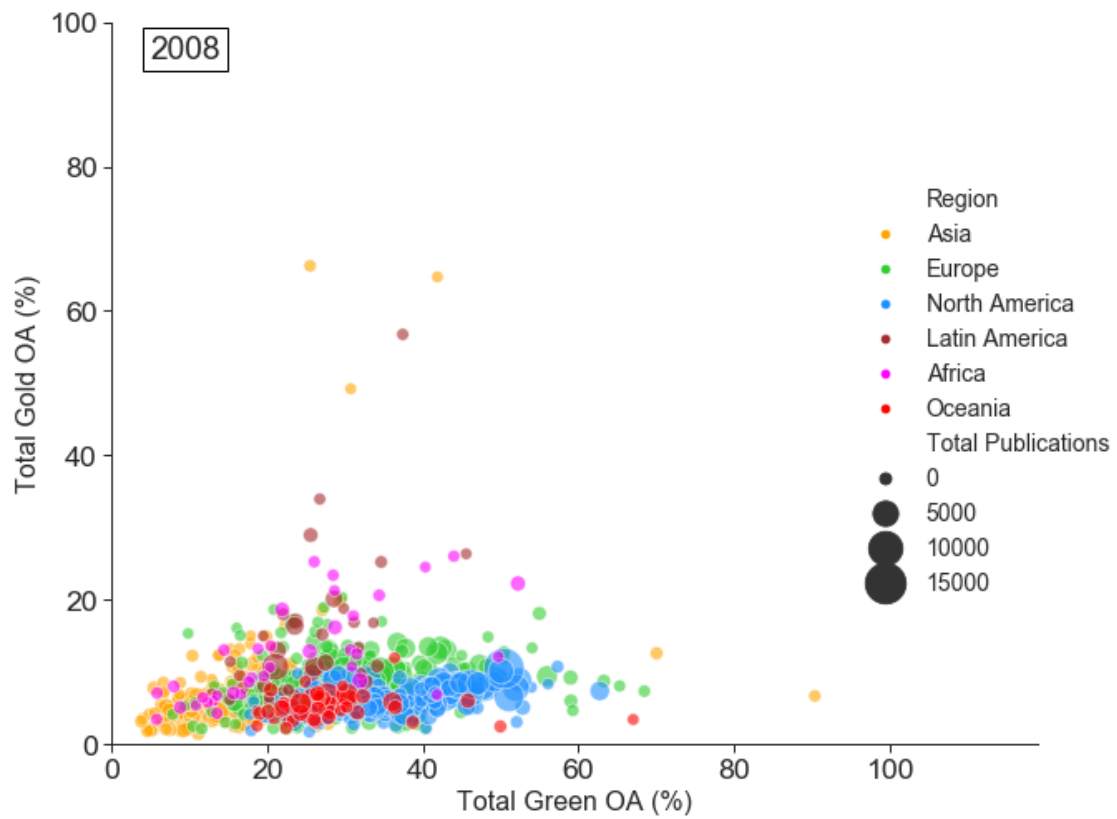
Overall universities in Oceania (Australia and New Zealand) and North America (Canada and the US) lag behind comparators in Europe (on repository-mediated open access) and Latin America (on open access publishing). Asian universities are highly diverse. As seen in Supplementary file 2, there are some high performers in the top 100s, particularly for open access publishing, but many also lag behind. Africa is also highly diverse but with a skew towards high performance, with an emphasis on open access publishing. This may reflect our sampling which is skewed towards institutions with the largest (formally recorded) publishing volumes, many of which receive significant portions of their funding from international donors with strong open access requirements. Latin American institutions show high levels of open access publishing throughout the period illustrated. This is due to substantial infrastructure investments in systems like SciELO starting in the 1990s.

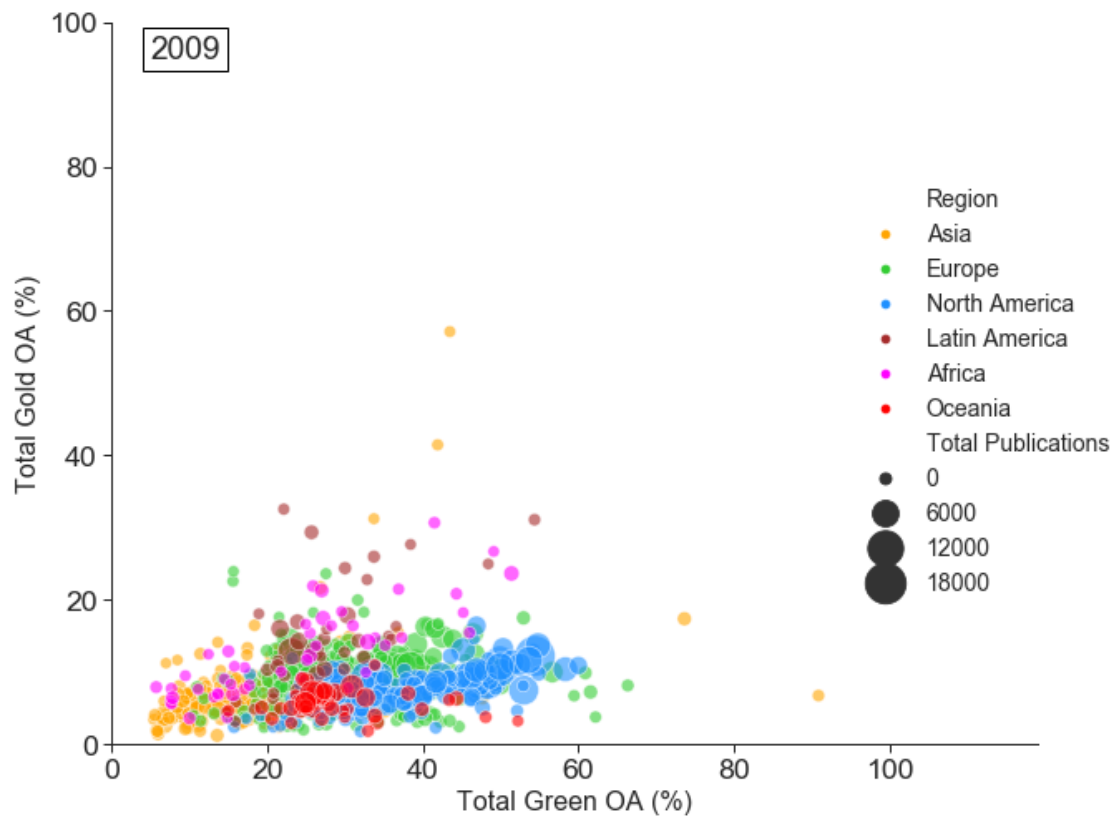
**Figure 3: Open access publishing (gold OA) vs repository-mediated open access (green OA) by institution for 2017 (and 2007-2018 for animated version). Each point plotted is a university, with size indicating the number of outputs analysed and colour showing the region. Articles can be open access through both publishing and repository routes so x and y values do not sum to give total open access.**

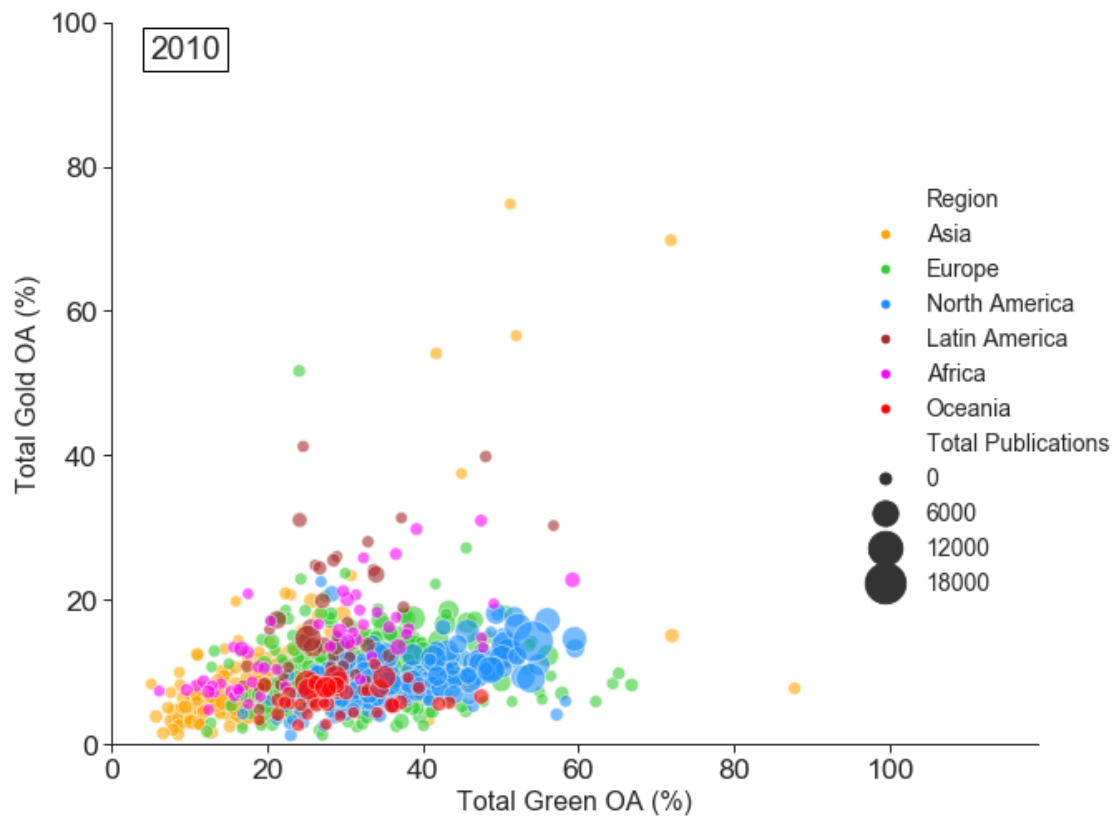


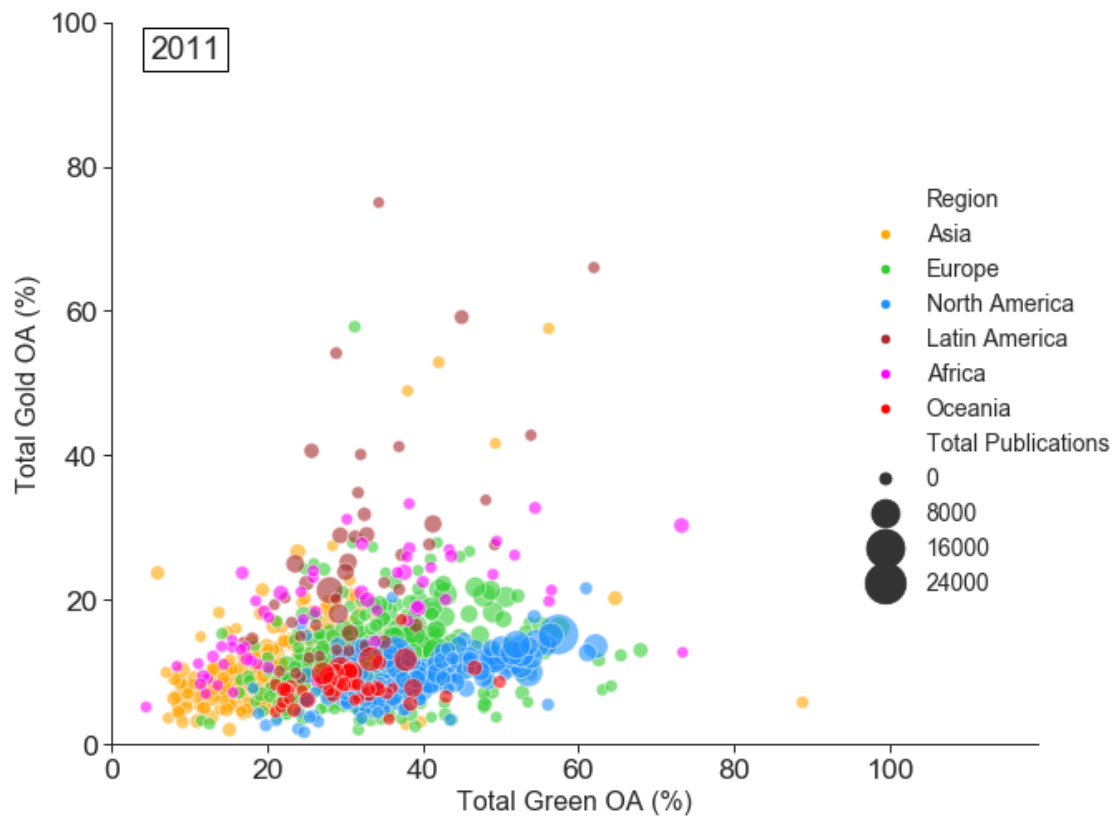
**Figure 3 - figure supplement 1**



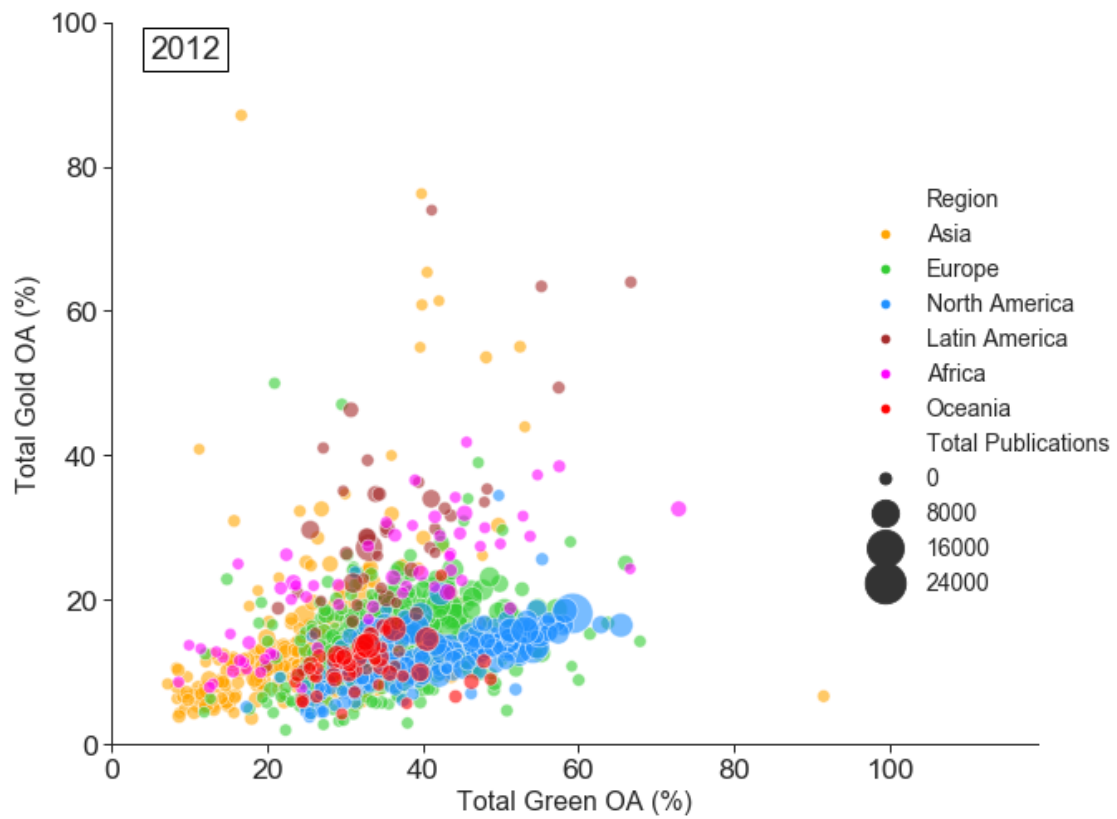


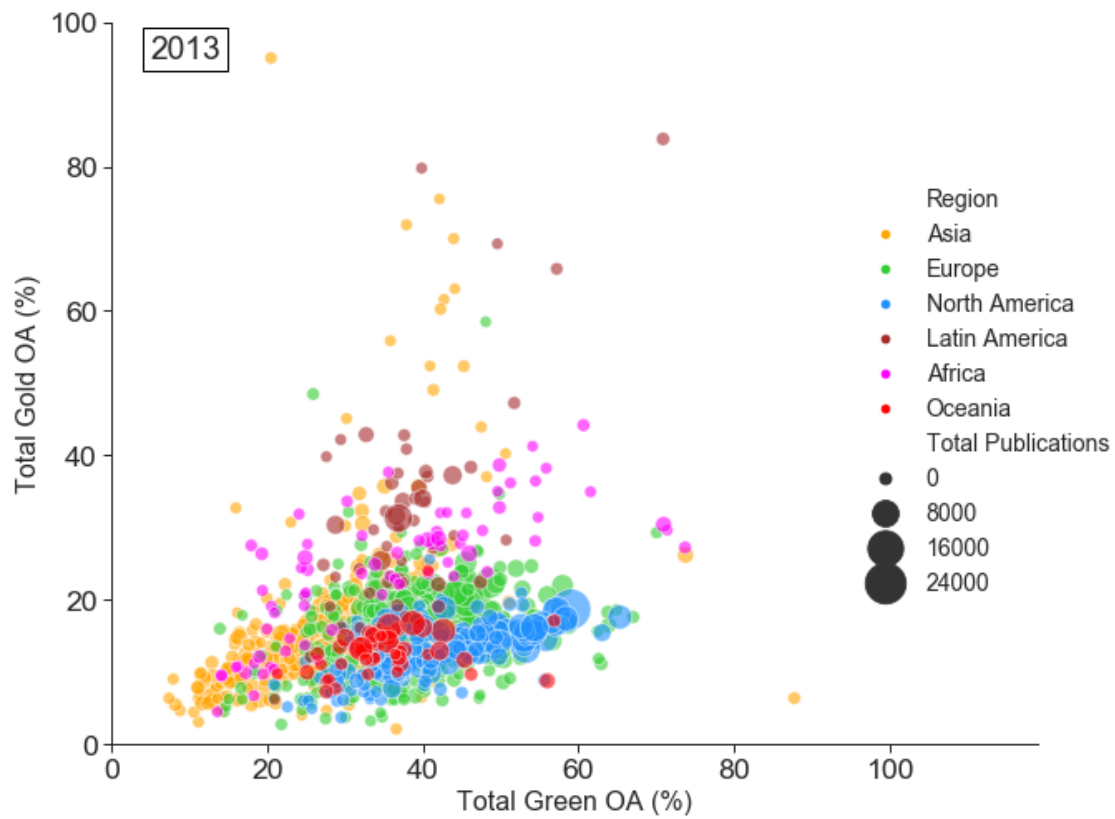


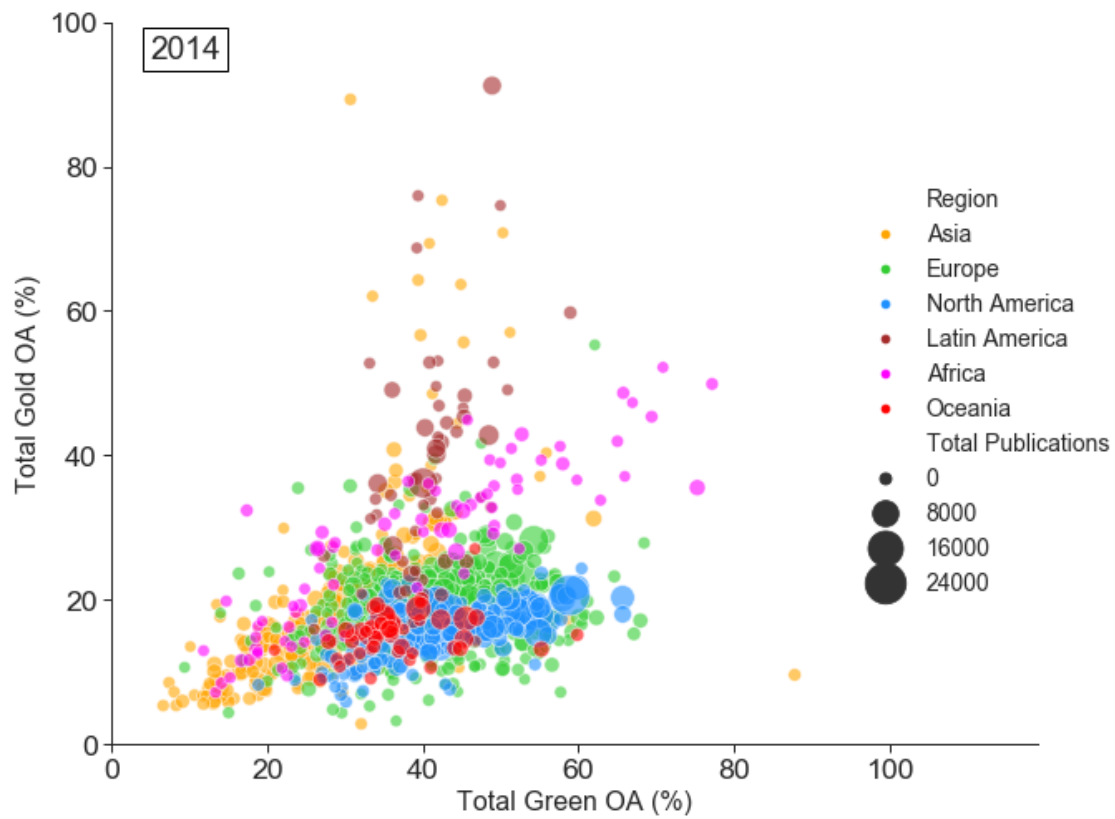


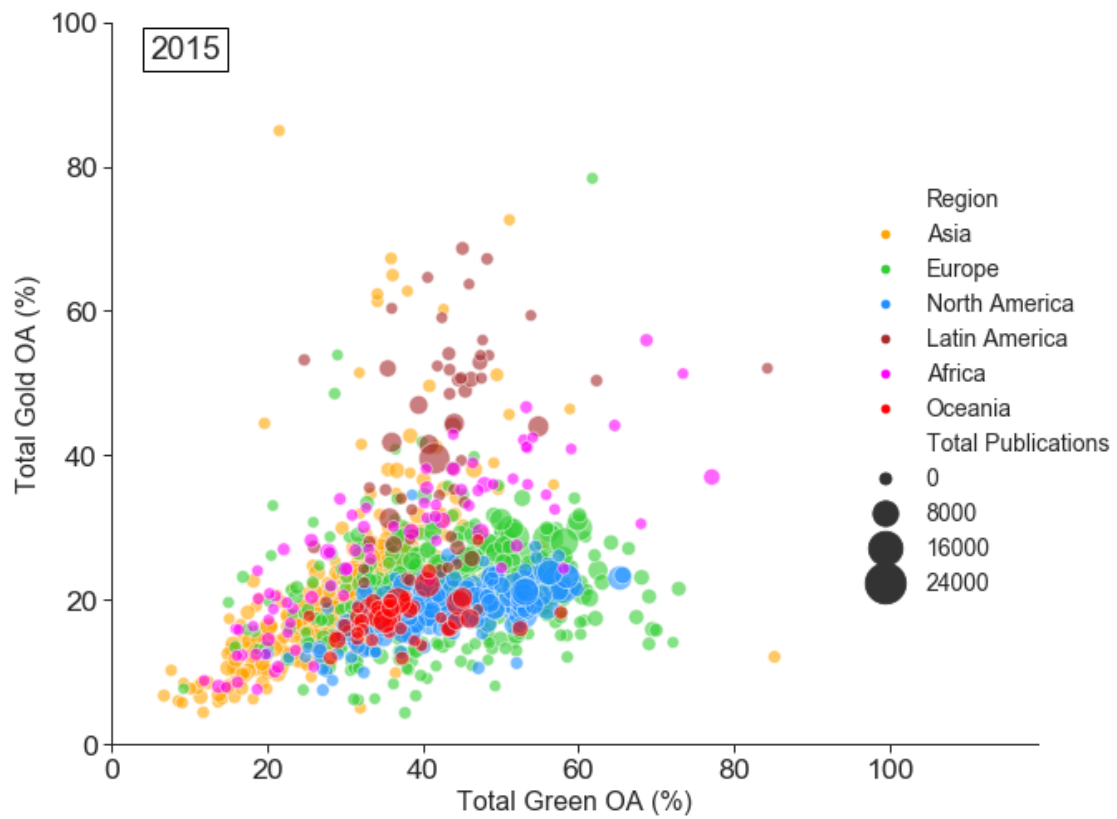


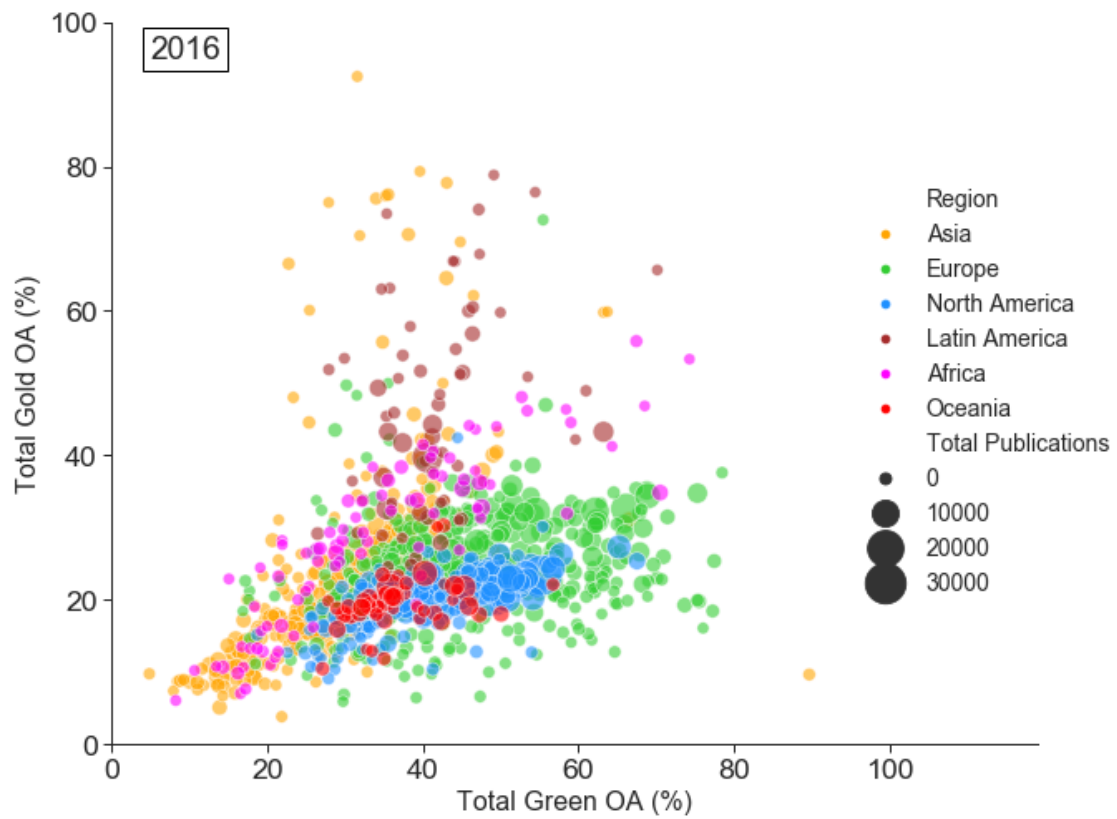


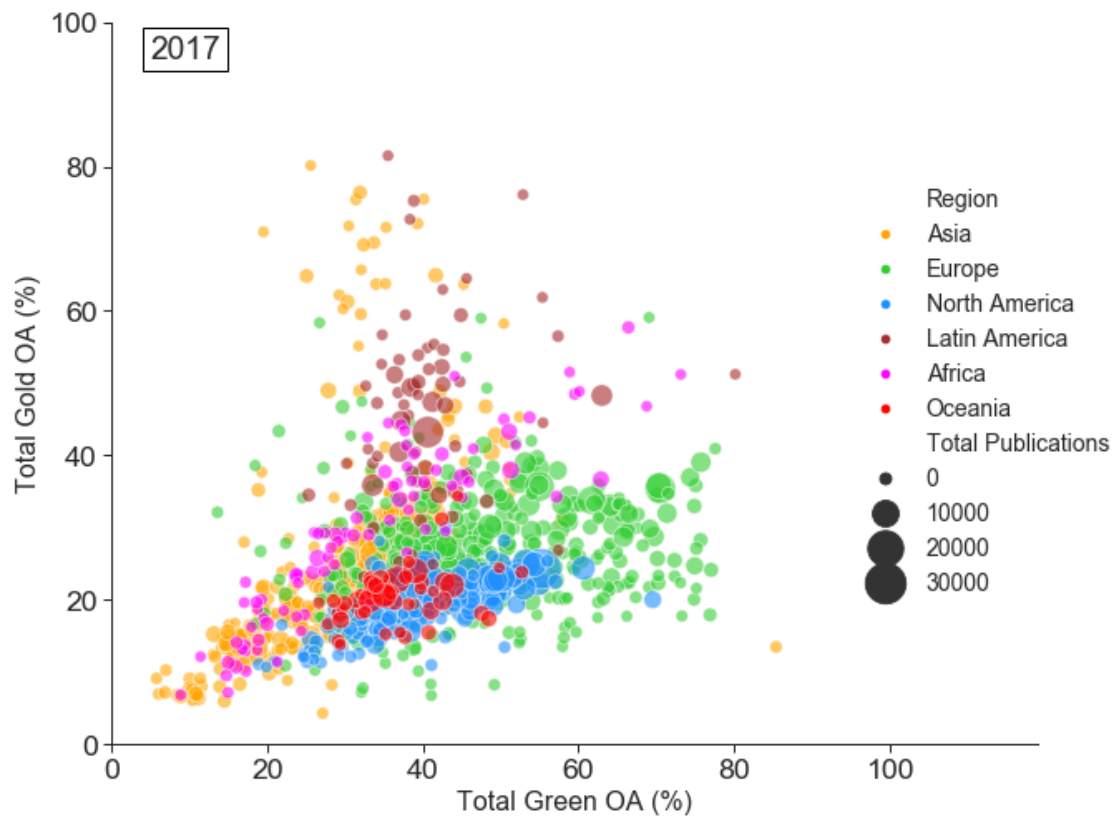


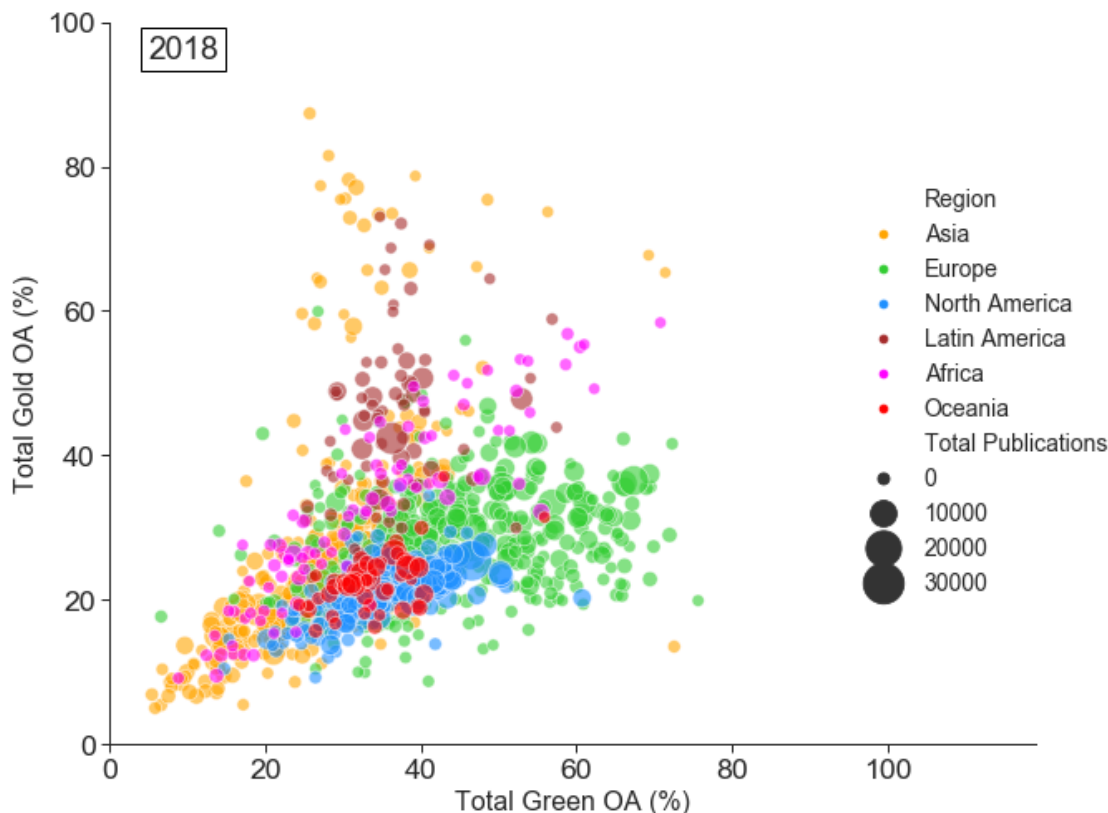












### 4.3 The effects of policy interventions

If our goal is to provide data on the effectiveness of interventions then our analysis should be capable of identifying potential effects of policy change. In 2012 the UK Research Councils, following the Finch Report, provided additional funding to individual universities to support open access publishing. The amount of additional funding relates to existing research council funding. In Figure 4a we show the annual change in open access publishing for three UK universities with the largest additional funding and three with significantly less additional funding (Lawson, 2018). In either 2012 or 2013 we observe a consistent increase in open access publishing across all the universities. As the additional funding tails off in 2015 the rate of growth falls back. Similarly, there are increases in proportions of hybrid open access which largely require article processing charges, and change in proportions of hybrid open access within all publisher-mediated open access (see Figure 4 - figure supplement 1a and 1b).

Figure 4b shows the growth of content in UK university repositories from 2000-2017, compared to two universities from other regions. In 2015 deposit of a research output in a repository became a requirement for eligibility for including in the UK Research Excellence Framework. This policy shift was profound because it relates to an assessment exercise and funding which covers all disciplinary areas and all universities. The dominance of the top 100 for both overall open access and repository-mediated open access by UK universities as well as the approach to 100% coverage being made by such a large number of universities is potentially driven in large part by that intervention.

Figure 4c focuses on the take-up of hybrid open access publishing options in the Netherlands as possible effects following deals with Springer in 2015, and Wiley in 2016. Across the Netherlands levels of publishing in hybrid open access journals show a sharp turn of increase from 2014 onwards with a less pronounced effects (more smooth increases) for publishing in pure open access (compare Figure 4 - figure supplement 1c and 1d).

Finally, in Figure 4d we show the possible effect of subtle differences in policy relating to acceptable embargo periods. UK Research and Funding Council policies have been aggressive in reducing embargo lengths mandating six months for STEM subjects and twelve months for HSS subjects. The effect of embargoes can be seen in data for repository mediated open access as a dip in the most recent years of publication. Using Unpaywall data from late 2019 we see a dip in repository-mediated open access performance for UK universities in 2018 but a limited effect on 2017. By comparison with three of the highest performing US universities, comparable in size and overall ranking (see Figure 2), we see an extended dip in performance, indicative of an acceptance of longer embargoes.



Figure 4: Monitoring the effect of policy interventions for selected groups of universities. Panel A shows the annual change in percentage (rolling current year percentage minus the previous year percentage) of gold OA for six UK universities. The first three universities are those with larger additional funding in contrast to the last three universities who received less additional funding. Panel B shows the annual percentage of green OA through the home institutional repositories of four UK universities compared to high performing universities from elsewhere. Panel C shows the annual percentages of hybrid OA at five universities in the Netherlands. Panel D shows three pairs of UK and US universities, selected based on having a similar size and level of green OA. The annual percentages of total green OA are depicted for each university.

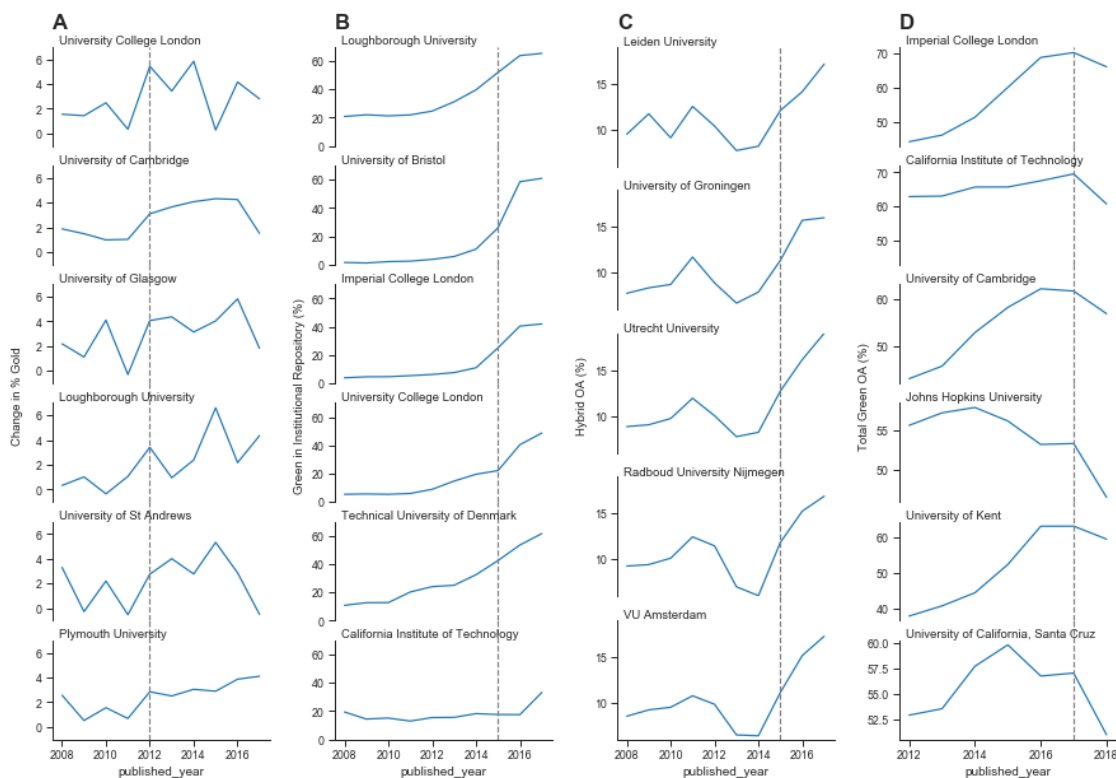
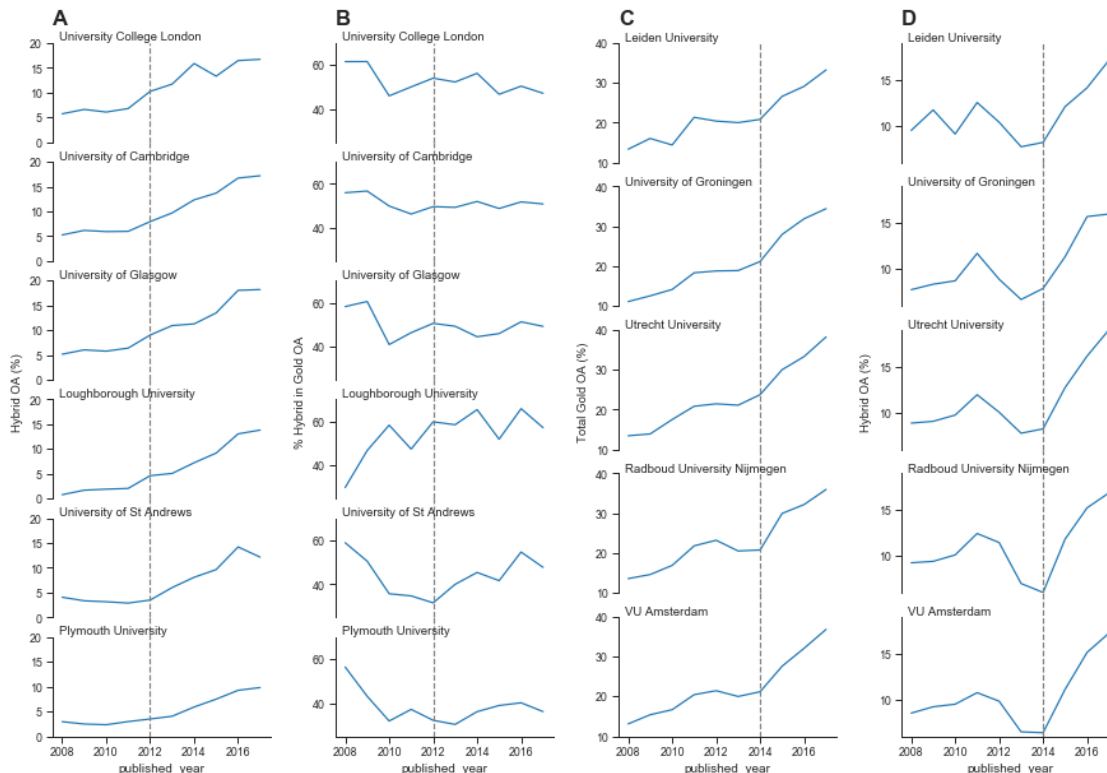


Figure 4 - figure supplement 1

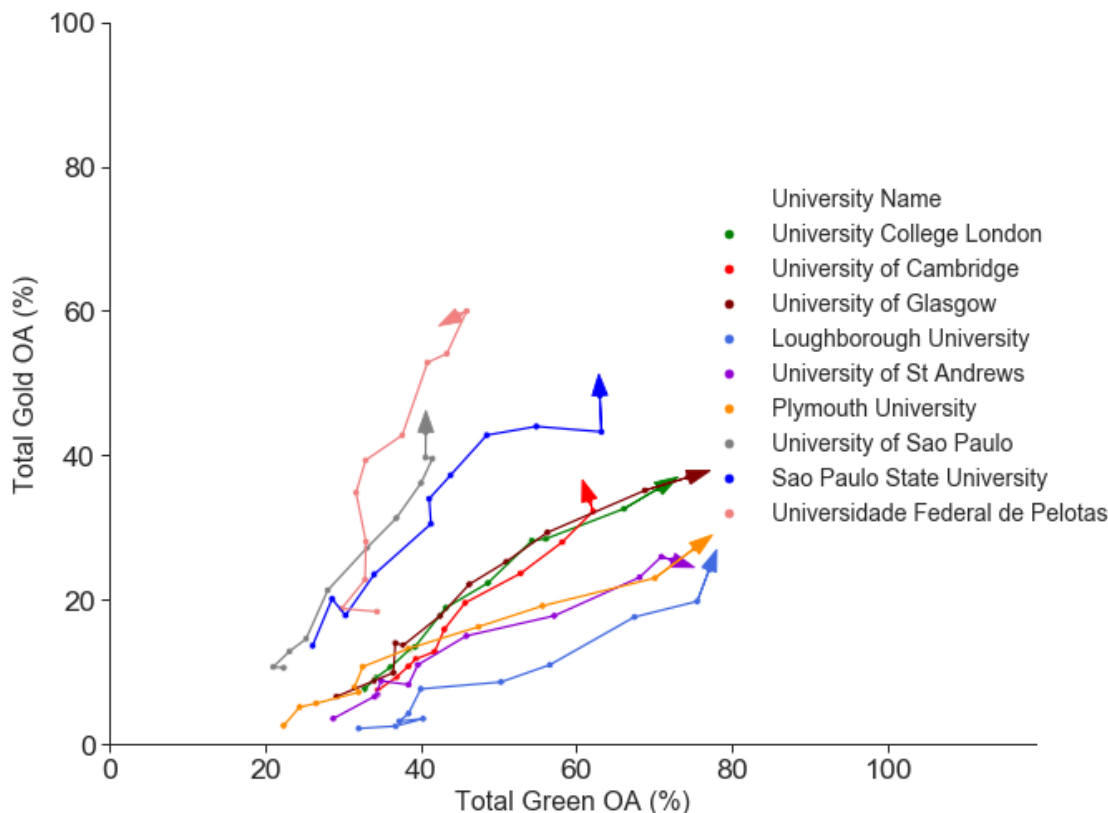


#### 4.4 Different institutional paths towards open access

In Figures 3 and 4 we see evidence of different paths towards open access, depending on the context and resources. The idea of mapping these paths is shown explicitly for a subset of universities in Figure 5. This shows the paths taken by two sets of UK universities and a selection of Latin American institutions over time. For UK universities that received substantial funding from the UK research councils for open access publishing three examples are shown. An alternate route, emphasising repository-mediated open access, is seen for three universities that received less additional funding.

In contrast the Latin American institutions already have high levels of open access publishing at our earliest time point, as discussed earlier. However, our data suggests a fall in overall open access amongst Latin American universities from 2012 onwards which we ascribe to an increased pressure to publish in “international” journals which are often subscription based, and for which Latin American scholars are reluctant or unable to pay hybrid article processing charges.

**Figure 5: Comparing different paths to open access (gold OA versus green OA ) for a selected set of universities from 2007 to 2017. This figure compares the different open access routes taken by three groups of universities: three UK universities that received substantial funding for open access publishing (combined gold and green OA increases), three UK universities that received less funding (more green OA focused), and three Latin American universities (more gold OA focused).**



## 5 Discussion

Our analysis of open access performance by research-intensive universities highlights the importance of robust policy and support in driving change. Geographies where there is a long history of infrastructure provision, such as Latin America, show very high levels of open access publishing. The United Kingdom is a particular case where there is consistently very high levels of open access, particularly that provided through repositories, in response to a strong and well supported policy environment. We also see that different institutions may choose to take different paths to delivering open access depending on resources, culture and systems in place.

Our results have significant implications for the details of policy interventions. Firstly, we have demonstrated the ability to detect signals of policy interventions in the behaviour of institutions. We see clear effects and results arising from the efforts of national funders and policy makers, particularly in the United Kingdom. The combined policy change and funding provided by the UK Research Councils in 2012 is associated with a increase in the level of open access publishing, and the level of increase appears to be associated with the level of funding provided. Similarly the requirement for outputs to be deposited in a repository for eligibility for the 2021 Research Excellence Framework is associated with substantial increase in repository-mediated open access around 2015.

Our results also may have implications for deciding on the effectiveness of directly funding open access publishing. It is perhaps surprising to some readers that the overall levels of open access

publishing in the UK are not higher. Specific funders, most notably the Wellcome Trust, have achieved very high levels of open access for articles from research they support through the provision of funding for open access publication. In addition the UK Research Councils invested significant resources in supporting gold open access. However, these have not translated to high levels of open access publishing across the full diversity of outputs of UK institutions. The majority gains over the past five years have come from repository-mediated open access.

In the animated version of Figure 3 (or the version over several years in Supplementary Figures) there is a clear signal of saturation with respect to open access publishing (gold open access) for European and North American universities. With few exceptions, institutions do not achieve levels of gold open access greater than 40% and this level is stable from 2014-2018. Similarly in Figure 4 we see evidence of shifts in response to stimuli (funding and policy interventions) which then stabilise. Even those UK universities with very high levels of repository open access (green) see a slowing down of the rise in levels a few years after the Research Excellence Framework policy intervention.

These signals suggest that achieving 100% open access may be very difficult, and possibly expensive to achieve. There will always be areas and cases where open access is challenging. Current challenges include disciplinary areas such as the humanities where suitable business models and venues are still developing, as well as types of content, particularly books where there is greater overlap between scholarly and general publishing. Both of these could be addressed rapidly through direct funding, but this may not be the most efficient approach. Achieving “100% open access” may therefore require a tighter definition of exactly which outputs are in scope. For those areas where we see signals of saturation much lower than 100% these are likely signals of the complexity of the system, and of large categories of outputs where open access is harder to achieve, or the motivation of institutions (including authors, libraries, and other support staff) to achieve it is lower.

Alongside this, the continued leadership of Latin American institutions on open access publishing levels is the continuation of a trend set more than a decade ago through the provision of publishing infrastructures. Taken alongside the clear response in the Netherlands for hybrid open access in response to publish and read agreements this suggests that increasing levels of open access publishing through article processing charges is potentially expensive compared to the costs of providing infrastructure.

Another interesting natural experiment is how the strength of funder actions is associated with overall change in levels of open access. In the Netherlands and the UK in particular, but also in the US, where funder policies have moved from encouragement, to mandates, to monitoring with sanctions for non-compliance there are substantial shifts in overall levels of open access. By contrast, in countries where policy remains effectively at the level of a recommendation, such as Australia, levels of open access lag significantly. Recent increases in reporting requirements by Australian funders might therefore be expected to lead to a detectable signal over the next 12-24 months.

Perhaps most interestingly in light of the debates surrounding Plan S is the evidence that it is universities in Latin America and Africa where levels of open access publishing are at their highest. As noted above, in Latin America this is a strong signal of the effectiveness of infrastructures such as SciELO in supporting uptake of open access practices. In the case of Africa there may be effects of funder requirements (with funders such as the Bill and Melinda Gates Foundation and Wellcome Trust that have strong open access requirements playing a significant role) as well as disciplinary spread. In both cases we are likely to have a limited view of the full diversity of research outputs

due to their poor capture in information systems from the North Atlantic.

In comparison to results presented by Robinson-Garcia et al. (2019, 2020), our analysis shows significantly improved performance of Latin American universities. This is driven by our inclusion of data sources beyond Web of Science, which has resulted in increased open access publishing for Latin America (Huang et al., 2020b). Similarly, our results also depict more clear observations of some high performing African and Asian universities. These produce a slightly different view in contrast to the one where UK universities dominate the top ranking positions for open access.

The value of analysis at the level of universities is that we gain a picture of open access performance across a diverse research ecosystem. We see differences across countries and regions, and differences between universities within countries. Overall we see that there are multiple different paths towards improving access, and that different paths may be more or less appropriate in different contexts. Most importantly, while further research is needed to unpick the details of the differences in open access provision, we hope this work provides a framework for enabling that longitudinal analysis to be taken forward and used wherever it is needed.

## 6 Acknowledgements

This work was funded by the Research Office of Curtin University through a strategic grant, the Curtin University Faculty of Humanities, and the School of Media, Creative Arts and Social Inquiry.

## 7 References

1. Aguado-López, E., & Becerril-García, A. (2019, November 6). Latin America’s longstanding open access ecosystem could be undermined by proposals from the Global North [LSE Latin America and Caribbean]. <https://blogs.lse.ac.uk/latamcaribbean/2019/11/06/latin-americas-longstanding-open-access-ecosystem-could-be-undermined-by-proposals-from-the-global-north/>
2. Agustini, B., & Berk, M. (2019). The open access mandate: Be careful what you wish for. *Australian & New Zealand Journal of Psychiatry*, 53(11), 1044–1046. <https://doi.org/10.1177/0004867419864436>
3. Barbour, G., & Nicholls, S. (2019). Open Access: Should one model ever fit all? *Australian Quarterly*, 90(3), 3–9. <https://www.jstor.org/stable/26687171>
4. Brainard, J. (2019). Scientific societies worry about threat from Plan S. *Science*, 363(6425), 332–333. <https://doi.org/10.1126/science.363.6425.332>
5. Debat, H., & Babini, D. (2019). Plan S in Latin America: A precautionary note. *PeerJ Preprints*, 7, e27834v2. <https://doi.org/10.7287/peerj.preprints.27834v2>
6. Gómez-Fernández, J. C. (2019). Plan S for publishing science in an open access way: Not everyone is likely to be happy. *Biophysical Reviews*, 11, 841–842. <https://doi.org/10.1007/s12551-019-00604-4>
7. Green, T. (2019). Is open access affordable? Why current models do not work and why we need internet-era transformation of scholarly communications. *Learned Publishing*, 32(1), 13–25. <https://doi.org/10.1002/leap.1219>

8. Haug, C. J. (2019). No Free Lunch—What Price Plan S for Scientific Publishing? *The New England Journal of Medicine*, 380, 1181–1185. <https://doi.org/10.1056/NEJMms1900864>
9. Huang, C.-K., Neylon, C., Brookes-Kenworthy, C., Hosking, R., Montgomery, L., Wilson, K., & Ozaygen, A. (2020a). Comparison of bibliographic data sources: Implications for the robustness of university rankings. *Quantitative Science Studies*, 1(2), 445–478. [https://doi.org/10.1162/qss\\_a\\_00031](https://doi.org/10.1162/qss_a_00031)
10. Huang, C.-K., Neylon, C., Hosking, R., Montgomery, L., Wilson, K., Ozaygen, A., & Brookes-Kenworthy, C. (2020b). Evaluating institutional open access performance: Sensitivity analysis. <https://doi.org/10.5281/zenodo.3696857>
11. Huang, C.-K., Neylon, C., Hosking, R., Montgomery, L., Wilson, K., Ozaygen, A., & Brookes-Kenworthy, C. (2020c). Data and Intermediate Queries for: Evaluating institutional open access performance: Methodology, challenges and assessment. <https://doi.org/10.5281/zenodo.3693221>
12. Huang, C.-K., Neylon, C., Hosking, R., Brookes-Kenworthy, C., Montgomery, L., Wilson, K., & Ozaygen, A. (2020d). Jupyter Notebooks for the article: Evaluating institutional open access performance: Methodology, challenges and assessment. <https://doi.org/10.5281/zenodo.3716063>
13. Iyandemeye, J., & Thomas, M. P. (2019). Low income countries have the highest percentages of open access publication: A systematic computational analysis of the biomedical literature. *PLOS One*, 14(7), e0220229. <https://doi.org/10.1371/journal.pone.0220229>
14. Larivière, V., & Sugimoto, C. R. (2018). Do authors comply when funders enforce open access to research? *Nature*, 562, 483–486. <https://doi.org/10.1038/d41586-018-07101-w>
15. Lawson, S. (2018). RCUK open access block grant allocation 2013-18. Figshare. [https://figshare.com/articles/RCUK\\_open\\_access\\_block\\_grant\\_allocation\\_2013-17/4047315](https://figshare.com/articles/RCUK_open_access_block_grant_allocation_2013-17/4047315)
16. McNutt, M. (2019). Opinion: “Plan S” falls short for society publishers—And for the researchers they serve. *Proceedings of the National Academy of Sciences*, 116(7), 2400–2403. <https://doi.org/10.1073/pnas.1900359116>
17. Mongeon, P., & Paul-Hus, A. (2016). The journal coverage of Web of Science and Scopus: A comparative analysis. *Scientometrics*, 106(1), 213–228. <https://doi.org/10.1007/s11192-015-1765-5>
18. Montgomery, L., Hartley, J., Neylon, C., Gillies, M., Gray, E., Herrmann-Pillath, C., Huang, C.-K. (Karl), Leach, J., Potts, J., Ren, X., Skinner, K., Sugimoto, C. R., & Wilson, K. (2018). Open Knowledge Institutions: Reinventing Universities. MIT Press Work in Progress. <https://wip.mitpress.mit.edu/oki>
19. Packer, A. L. (2009). The SciELO Open Access: A Gold Way from the South. *Canadian Journal of Higher Education*, 39(3), 111–126. <http://journals.sfu.ca/cjhe/index.php/cjhe/article/view/479>
20. Piwowar, H., Priem, J., Larivière, V., Alperin, J. P., Matthias, L., Norlander, B., Farley, A., West, J., & Haustein, S. (2018). The state of OA: a large-scale analysis of the prevalence and impact of Open Access articles. *PeerJ*, e4375. <https://doi.org/10.7717/peerj.4375>

21. Piwowar, H., Priem, J., & Orr, R. (2019). The Future of OA: A large-scale analysis projecting Open Access publication and readership. *BioRxiv*. <https://doi.org/10.1101/795310>
22. Rabesandratana, T. (2019). The world debates open-access mandates. *Science*, 363(6422), 11–12. <https://doi.org/10.1126/science.363.6422.11>
23. Robinson-Garcia, N., Costas, R., & van Leeuwen, T. N. (2019). Indicators of Open Access for universities. *ArXiv*, 1906.03840. <https://arxiv.org/abs/1906.03840>
24. Robinson-Garcia, N., Costas, R., & van Leeuwen, T. N. (2020). State of Open Access penetration in universities worldwide. *ArXiv*, 2003.12273. <https://arxiv.org/abs/2003.12273>
25. Siler, K., Haustein, S., Smith, E., Larivière, V., & Alperin, J. P. (2018). Authorial and institutional stratification in open access publishing: The case of global health research. *PeerJ*, 6, e4269. <https://doi.org/10.7717/peerj.4269>
26. Sinha, A., Shen, Z., Song, Y., Ma, H., Eide, D., Hsu, B.-J., & Wang, K. (2015). An Overview of Microsoft Academic Service (MAS) and Applications. In *Proceedings of the 24th International Conference on World Wide Web (WWW '15 Companion)*. ACM, New York, NY, USA, 243-246. <http://dx.doi.org/10.1145/2740908.2742839>
27. Van Noorden, R. (2019, May 15). Indonesia tops open-access publishing charts. *Nature News*. <http://doi.org/10.1038/d41586-019-01536-5>
28. Wang, X., Cui, Y., Xu, S., & Hu, Z. (2018). The state and evolution of Gold open access: A country and discipline level analysis. *Aslib Journal of Information Management*, 70(5), 573–584. <https://doi.org/10.1108/AJIM-02-2018-0023>
29. Wang, K., Shen, Z., Huang, C., Wu, C.-H., Eide, D., Dong, Y., Qian, J., Kanakia, A., Chen, A., & Rogahn, R. (2019). A Review of Microsoft Academic Services for Science of Science Studies. *Frontiers in Big Data*. <https://doi.org/10.3389/fdata.2019.00045>