

Open access levels: a quantitative exploration using Web of Science and oaDOI data

Jeroen Bosman  [0000-0001-5796-2727](https://orcid.org/0000-0001-5796-2727)

- Utrecht University Library, j.bosman@uu.nl

Bianca Kramer  [0000-0002-5965-6560](https://orcid.org/0000-0002-5965-6560)

- Utrecht University Library, b.m.r.kramer@uu.nl

ABSTRACT

Across the world there is growing interest in open access publishing among researchers, institutions, funders and publishers alike. It is assumed that open access levels are growing, but hitherto the exact levels and patterns of open access have been hard to determine and detailed quantitative studies are scarce. Using newly available open access status data from oaDOI in Web of Science we are now able to explore year-on-year open access levels across research fields, languages, countries, institutions, funders and topics, and try to relate the resulting patterns to disciplinary, national and institutional contexts. With data from the oaDOI API we also look at the detailed breakdown of open access by types of gold open access (pure gold, hybrid and bronze), using universities in the Netherlands as an example. There is huge diversity in open access levels on all dimensions, with unexpected levels for e.g. Portuguese as language, Astronomy & Astrophysics as research field, countries like Tanzania, Peru and Latvia, and Zika as topic. We explore methodological issues and offer suggestions to improve conditions for tracking open access status of research output. Finally, we suggest potential future applications for research and policy development. We have shared all data and code openly.

KEYWORDS

open access, Web of Science, oaDOI, green open access, hybrid open access, gold open access, bronze open access, science policy, publishing, scholarly communication

INTRODUCTION

Open access is an important policy objective for many countries, research funders and institutions, often with higher goals: increasing speed, collaboration and innovation in research. This in turn can help solve urgent societal issues and questions. There is heated debate among stakeholders on how to increase open access. However, quantitative empirical studies are relatively scarce and reports on open access levels are often contradictory. This is at least in part caused by the difficulty to generate granular data on open access levels. The introduction of those data in Web of Science in December 2017 is a potential big step forward in that respect. We explore those data, with 5 goals: (1) test the usability of the new open access data and methods, (2) explore and present data: show open access levels by research field, language, country, institution, funder and topic (3) suggest explanations and (by that) also suggest further and hypothesis testing research, (4) show applicable uses of this method for informing policy development and monitoring their effects and (5) suggest improvements in data availability and methods. We are very much aware of the discussion on the pros and cons of the various routes towards open access but judging on the viability, equitability and efficiency of those is not the focus of this article/

Earlier large scale attempts to determine OA levels

For many years determining the success of open access had been done by looking at the growth of the number of open access journals, the growth of content archived in institutional or subject repositories or the growth of the number of articles published in the larger open access journals. More comprehensive and granular analysis were difficult because of lacking data. Large bibliographical databases (such as Scopus and Web of Science) lacked and sometimes still lack open access filters or used crude filters that only recognized papers that appeared in full gold open access journals, thus missing hybrid open access and green open access. Articles shared through green open access have been harvested by search engines such as BASE search, but that does not allow combining the information with information on the gold open access status of the same paper.

Despite these limitations there have been some good attempts to answer the question of the level of open access availability of journal articles. The broadest and most systematic approach has been that by ScienceMetrix, a Canadian science data analytics company. In a study commissioned by the European Union (Archambault et al. 2014) they used large samples (245K and 1M) of articles published from 1996-2013 indexed in Scopus and matched that with a self-built database of gold and green open access articles.

ScienceMetrix found high aggregate levels of open access, by using a harvester on Google on Google Scholar and also by taking all types of open access into account, including papers shared through academic social networks such as ResearchGate and on researcher's own websites. That way they found global OA levels over 45% for 1996-2013. ScienceMetrix was also able to present breakdowns by country and field. ScienceMetrix later moved to using WoS as baseline reference for the total number of papers in data and publications that still form the basis for policy evaluations for the European Research Area by the European Union (ScienceMetrix 2016). The very high levels of OA, especially green, found remain, probably largely caused by open versions of papers shared by authors

themselves on sharing platforms such as ResearchGate, that have recently become a target of legal action by publishers because of alleged copyright infringements. An idiosyncrasy in the analyses of ScienceMetrix is that they classify papers deposited in PubMed Central as gold, not green OA, with the argument that deposition is in large part done by publishers, not individual authors.

More recently Piwowar et al. (2017), also using their own database of open access articles information (oaDOI) tried to increase the reliability by using three different 100K samples (from WoS, Crossref and Unpaywall usage) to conclude that *at least* 28% of the literature is open access, with levels to to 45% for 2015. Notably Piwowar et al. were also able to determine types of green only and gold open access, confirming the substantial amounts of bronze open access that Laakso & Bjork (2013) had found earlier. However that study did determine levels by publisher and field, but not by country, language, institution, funder and topic.

Finally, at the national level there are notable studies from the UK and Austria. In the UK quantitative studies are those commissioned by Universities UK (RIN 2015, using samples of 5K and 10K articles, and Research Information 2017) that specifically focussed on (potential) immediate OA and additional OA in the first 2 years after publication. In Austria the Austrian Science fund leased data on the open access status of their output (Kunzmann & Reckling 2017).

METHODOLOGY

Data availability

oaDOI was launched in 2016 by the team behind [ImpactStory](#) (Heather Piwowar and Jason Priem). Inspired by similar initiatives as [DOAI](#) and the [Open Access Button](#), it was introduced as a service to check open access availability of journal articles identified by their DOI. oaDOI uses information on journal open access status, article license information and availability of versions of an article in disciplinary and institutional repositories to provide detailed information on open access at the article level, including hybrid open access. oaDOI offers access through their data through various ways, including the [Unpaywall](#) browser plugin and an [open API](#).

In December 2017, [Web of Science](#) (WoS), a pay-walled citation database supplied by Clarivate Analytics, started including information on open access availability of publications using article-level information from oaDOI. Previously, information on OA availability on publications in Web of Science was based on journal level, only including gold Open Access journals indexed in DOAJ.

Using oaDOI data, Web of Science currently labels open access articles as 'Gold', 'Green Accepted' or 'Green Published' (Fig. 1). Articles labeled as gold open access in Web of Science are all articles that are detected as freely available from the publisher. No distinction is made between articles in gold open access journals, hybrid journals or subscription

journals that make (some) articles freely available, but without a license for re-use (read-only), often after a certain amount of time (delayed open access). In this paper, we refer to this entire group as 'all gold' open access.



Figure 1 Example of Open Access display in Web of Science.

Articles labeled as 'Green Accepted' or 'Green Published' in Web of Science are all articles for which either the *accepted version* (manuscript after peer review, but without publisher formatting) or the *published version* (with publisher formatting) can be retrieved from a repository harvested by oaDOI. *Submitted versions* (author manuscript before peer review) from the same repositories are not included, even though this information is collected by oaDOI. This excludes, for example, much of the content on ArXiv, which is included as a data source in oaDOI. oaDOI does not include open access (either preprints, peer-reviewed manuscripts or publisher versions) through academic networks sites such as ResearchGate and Academia.

When multiple open access versions of an article are available, oaDOI prioritizes publisher-hosted content (i.e. gold, hybrid or bronze), then versions closer to the version of record (i.e. for green OA, published version over accepted version over submitted version). Web of Science only includes the 'best' OA location as determined by this algorithm. The implication of this is that levels of green OA reported in Web of Science only indicate *additional* green OA, that is, publications not also available as Gold OA. Hence below we refer to this as 'green only'.

Through the oaDOI API, more detailed information about the open access status of individual articles can be obtained. Using information about journal open access status and available licenses for individual articles, the 'all gold' category in Web of Science can be broken down into the following subcategories (descriptions taken from Piwowar et al., 2017):

- Gold: Published in an open-access journal (as defined by inclusion in the [Directory of Open Access Journals \(DOAJ\)](#)). Henceforth, we refer to this as 'pure gold'.
- Hybrid: Free under an open license in a toll-access journal
- Bronze: Free to read on the publisher page, but without an open license

Clarivate itself has shared some information on the open access information in Web of Science (Clarivate Analytics 2017a and 2017b).

The detection of green only OA by oaDOI has undergone recent changes: while previously, BASE was used as the primary source of green OA detection, oadOI is currently harvesting repositories themselves. Sources covered include disciplinary repositories as PubMed Central and ArXiv, and institutional repositories. oaDOI states it currently harvests thousands of institutional repositories (Impactstory 2017a) and is actively adding more. At the time of writing, it was not clear how complete the coverage of institutional repositories was, e.g. how many of the currently 3252 active repositories in OpenDOAR and/or the 6046 sources in BASE are covered (Directory of Open Access Repositories 2018, Bielefeld University Library 2018). It is also not known whether there might be some systematic biases as a result of yet incomplete coverage, e.g. in favour of larger repositories, or US and European ones. Spot checks reveal that of the green only OA included in Web of Science, relatively little is coming from institutional repositories, and most content is retrieved from PubMed Central.

Another recent change at oaDOI is the method used for detection of green only OA versions (*published*, *accepted* and *submitted* versions). While oaDOI previously used OAI-PMH metadata to detect the version of a green OA article (Impactstory 2017b), the oaDOI team now relies on an automated, heuristic-based approach, looking for CrossMark, publisher watermarks etc. on PDF-files hosted in institutional repositories (Priem 2017a). Currently, little is known about the accuracy of both methods.

According to their documentation, WoS updates its open access information weekly (Clarivate Analytics, 2017b). Information on pure gold and green only OA is also updated weekly in the oaDOI database. For bronze OA, it was confirmed that detection takes longer, with a monthly or even yearly update schedule (Priem 2017b).

The availability of article-level open access data in Web of Science, in combination with existing search and filtering options, enables detailed comparisons of open access levels for different research fields, languages, countries, institutions, funders and topics. For a number of these parameters (research fields, languages, countries and institutions), Web of Science offers enhanced field indexing, enabling unambiguous identification of these concepts. For example, searching for 'Utrecht University' in the 'Organization Enhanced' field yields publications from Utrecht University regardless of naming variants used in the publication itself. For funders, there is a dedicated metadata field in Web of Science, but no enhanced indexing for this field. This means that variations of a funder name need to be searched or selected manually to be included. Topics can be searched by words or phrases in the topic field, which simultaneously searches the title, abstract and keywords of a publication.

A number of fields, including research fields, languages, countries, institutions and funders, are also available as filters to limit a given search. These filters each show the first 100 lemmas in the respective field (by record count), enabling selection based on frequency of occurrence.

Depending on one's institutional subscription, Web of Science offers access to a number of citation databases (Web of Science Core Collection). At the time of data collection (December 20-29 2017), our university's license covered the following databases:

- Science Citation Index Expanded (SCI) (1900-present)
- Social Sciences Citation Index (SSCI) (1900-present)
- Arts & Humanities Citation Index (A&HCI) (1975-present)
- Emerging Sources Citation Index (ESCI) (2015-present)

For the publication years 2015-2017, WoS Core Collection at our institution includes the Emerging Sources Citation Index with 718,473 records, covering 7155 journals at 20180106 according to the ESCI master journal list (Clarivate Analytics 2018) many of those described by Clarivate as “of regional importance and in emerging scientific fields”. ESCI represents over 10% of the WoS Core databases for those years. But still, WoS only captures a selection of the world scholarly publication output with underrepresentation of non-English and non-Western literature and having almost no representation of scholarly output published in books.

We also know the coverage of fully open access journals is not complete in Web of Science. A 2017 analysis of coverage of OA journals showed that of 24,417 OA journals identified in DOAJ, Directory of Open Access Scholarly Resources (ROAD), PubMed Central (PMC) and Open APC (OAPC), Web of Science covered only 1851. Of the 9,954 journals in DOAJ at the time this analysis was done (September 2017), only 1236 were included in Web of Science (Rimmert et al. 2017). It is not sure whether this analysis includes the ESCI part of Web of Science. In a post from Clarivate of one year earlier (Clarivate 2016a) it stated it had 930 OA journals in the SCI, 133 in the SSCI, 75 in the AH&I and added 1640 through ESCI to a reach a deduplicated total of 2744.

Data collection and analysis

Data were collected from Web of Science, between December 20, 2017 and January 5, 2018, using the full Web of Science Core Collection. Thus, the current study does not use a sample but the full WoS database. All searches were limited to the 12.3M articles and reviews from the publication years 2010-2017.

Our analysis focuses on research fields, languages, countries, institutions, funders and topics. Within each of these, the selection of units to analyze was based on either complete representation, frequency counts (provided by the filtering options in Web of Science) or more subjective selection, aimed at achieving diverse representation but also including cases of special interest. Where possible, the enhanced index fields from Web of Science were used to enable unambiguous identification. Detailed selection criteria for each variable are listed below.

- **Fields:** based on Web of Science field Research areas (field name: SO; enhanced-indexed).
For each main discipline group (Life Sciences & Biomedicine, Physical Sciences & Technology, Social Sciences, Arts & Humanities), fields were selected that were expected to have either high or low open access levels. We also took care to include the largest fields and have variety in terms of pure and applied science. Note that Web of Science

research areas are defined at the journal, not the article level and that journals can have been assigned to more than one research area.

- **Languages:** based on WoS filter Languages (enhanced-indexed).
Top 7 languages were selected based on frequency count. Spanish just from Spain-affiliated institutions was added as a separate unit.
- **Countries:** based on WoS field Country (field name: CU, enhanced-indexed).
For each continent, 10-15 countries were selected from WoS top-100 list based on frequency count. Further details per continent:
 - Africa: all countries with >5000 articles+reviews in total in 2010-2017
 - Asia: countries with >7000 articles+reviews in total in 2010-2017
 - Europe: EU-28 plus Norway and Switzerland
 - Latin America: all countries with >5000 articles+reviews in total in 2010-2017
 - Middle East and Eastern Europe: all countries with >7000 articles+reviews in total in 2010-2017
 - Worldwide: largest countries from North America and Oceania, supplemented with a subjective selection of countries based on regional representation and/or special interest regarding OA policies or practices.
- **Institutions:** based on WoS field Organization-enhanced (field name: OG, enhanced-indexed).
A set of mostly rich western universities was selected because of their vocal support for open access. Some others were added for comparison. We also subjectively selected institutions from the global south to which we added institutions from Moscow and China, with the World bank for comparison.
- **Funders:** based on WoS field Funding agency (field name: FO, not enhanced-indexed).
This first exploration was loosely based on frequency counts. We looked at the 100 most often mentioned funder names, grouped name variants together and selected 16 of those, including important funders from various countries around the world. The Gates Foundation was added for comparison. Funder information in Web of Sciences is derived from funding acknowledgement in the papers themselves and not normalized. Additional sources are Medline and ResearchFish. SCIE records have this funding information added since 2008 and SSCI records since March 2015 (Clarivate Analytics 2016b). For the year 2016 some 59% of the articles and reviews have funding information in the WoS record.
- **Topics:** based on WoS field Topic (field name: TO, not enhanced-indexed; searches title, abstract and keywords).
Selection of topics was done subjectively, including a number of the UN sustainable development goals (United Nations 2015).
- **Dutch universities for in-depth analysis:** based on WoS field Organization-enhanced (field name: OG, enhanced-indexed).
The initial selection was comprehensive, including all 14 Dutch universities. For longitudinal in-depth analysis, a selection of 4 universities was made: 2 general universities (Leiden and Utrecht), one technical university (Delft) and a mixed type (Wageningen).

For each unit of analysis, the total number of articles and reviews were recorded for each publication year between 2010 and 2017, as well as the corresponding number of open access articles and reviews. Numbers were collected manually in a spreadsheet and percentages open access calculated and plotted.

To investigate levels of pure gold, hybrid and bronze open access within the Web of Science 'all gold' open access category, sets of full records for each parameter were exported from Web of Science in tab-delimited Win (UTF-8) format. This was done for all Dutch universities for 2016, and for 4 selected universities for all the years from 2010-2017. Sets contained all records that were categorized as 'gold' open access in Web of Science.

Using an R-script (Kramer 2018), the DOIs from each set of records were extracted and used to query the oaDOI API. For each DOI, the following information was collected:

- whether an OA-version of the article was found
- whether an OA-version was found on the publisher website or in a repository
- whether the OA-version has an open license
- whether the journal the article is published in is included in DOAJ
- the URL where the OA-version of the article can be found

From the above, the number and percentage of gold (journal in DOAJ), hybrid (journal not in DOAJ, article with open license) and bronze (journal not in DOAJ, article without open license) articles is calculated for each set of records. The results are plotted together with the percentage of green only open access as retrieved from Web of Science directly.

It is important to realize that while the oaDOI API returns all locations where an open access version of an article is found, it also indicates the 'best' location - prioritizing publisher-hosted content over content found in repositories. For this analysis, only information on this 'best' OA version was used for each DOI.

To assess the relative contribution of different disciplines to the observed levels of pure gold, hybrid, bronze and green only open access, information from the full record export from Web of Science was used. Each record in Web of Science is assigned one or more research areas, taken from a standardized list of over 150 lemmas, classified into five broad categories (Clarivate Analytics 2017c):

- Arts & Humanities
- Life Sciences & Biomedicine
- Physical Sciences
- Social Sciences
- Technology

For each record analyzed, the first assigned research area was taken and mapped to one of the main categories (combining Physical Sciences and Technology into one category). This information was then combined with the observed open access status. The analysis was done for the sets of articles and reviews categorized as 'gold' open access in Web of Science from four Dutch universities, for two publication years (2014 and 2016).

All data (Kramer & Bosman 2018) and code (Kramer 2018) have been shared openly.

Caveats

Because of the way Web of Science and oaDOI are designed and function, there are some caveats we have to keep in mind while presenting and interpreting the results.

1. Green OA levels are potentially underreported for the following reasons:
 - All levels for green OA should be interpreted as 'green only' (i.e. not also available as gold OA), rather than as the total share of articles available as green OA.
 - WoS uses a relatively strict demarcation of green only OA: open access in the form of author submitted versions and preprints are not included. oaDOI also does not include availability from academic networks sites such as ResearchGate and Academia, as is done in some other assessments of OA.
 - Green open access data available to us via WoS or the oaDOI API may be incomplete because not all repositories are being harvested yet.
2. There may be a slight time lag in updating oaDOI information in Web of Science, either due to the frequency with which WoS checks oaDOI and/or the frequency with which the oaDOI database is updated. Both Clariivate and Impactstory state that their update frequency of OA data is weekly. Still that could mean in theory that a maximum of 2 weeks of OA data is lacking (up to 4% of a year).
3. Due to the inclusion of the Emerging Sources Citation Index (ESCI) in Web of Science in 2015, the absolute number of articles+reviews can be substantially higher for publication years 2015-2017. Depending on the level of OA of papers in the ESCI part of WoS, the effect can be markable, especially for countries in the global south.
4. Our analysis reflects the state of open access at one point in time (December 2017, with just a few analyses using January 2018 data). This means that time series of OA levels for the 2010-2017 years of publication presented in the results should not be simply interpreted as developments with 'growth' and 'decline'. OA levels found for all years are still subject to change, though most of that change will probably occur in the levels of the most recent publications years.
5. The limited and selective view of Web of Science on the total global scholarly publication output, makes that all our results indeed only describe a part of what is going on worldwide and there may be biases because of WoS' selection of journals, including open access journals.
6. For units of analysis that have a low absolute number of publications, observed changes in levels of OA may be less reliable and appear to be more volatile.

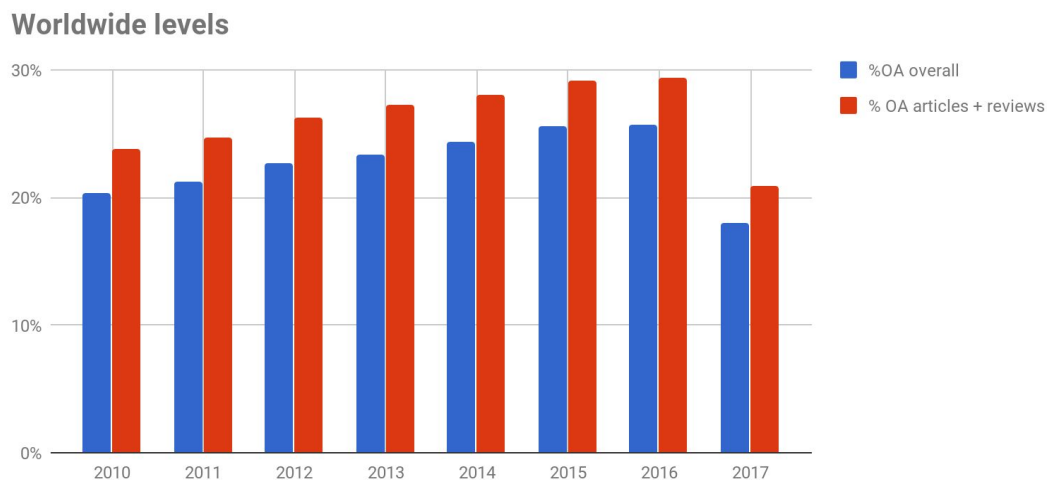
RESULTS

In the sections below we first present results of the exploration of open access levels of articles and reviews published between 2010 and 2017. There are seven sections in which we present worldwide results and results by field, language, country, institution, funder and topic. This is followed by a more in depth exploration of the levels of the different types of open access using a case study of Dutch universities.

Worldwide levels

Globally (as far as Web of Science can be said to capture the global picture) open access levels of articles and reviews are higher than those of all content, that also includes some 25% other publication types: meeting abstracts, letters, editorial content etc. (Fig. 2). Articles and reviews published in more recent years are more often open access, up to almost 30% in 2015 and 2016.

Figure 2 Open Access levels worldwide 2010-2017.



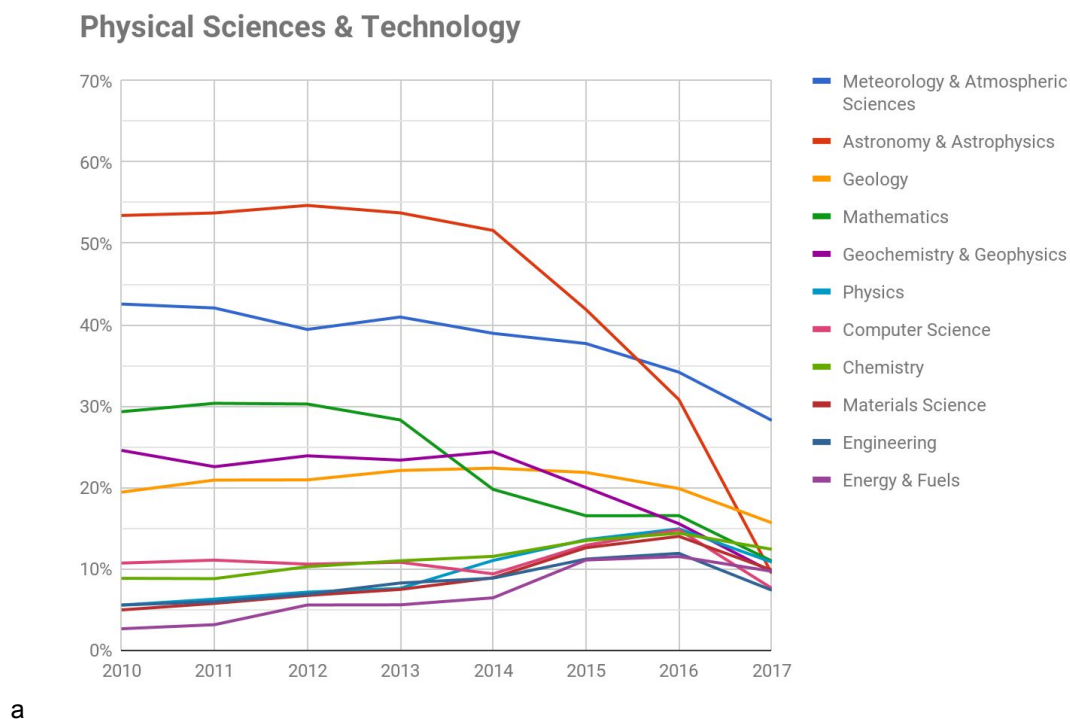
Most striking here is the low level for 2017. There are a few possible explanations:

- (1) Publishers only make some of their journal content openly available (bronze open access) after a number of years (mostly 1 or 2); we know from earlier research (Laakso & Björk 2013) that this involves many hundreds of journals and over 110 thousand articles in 2011. If that number has increased at the same rate as the overall output between 2010-2017 (up to 40%), then that delayed OA (part of bronze OA) could account for all of the 'drop' of OA levels for 2017 output.
- (2) There could be some lag effect due to the update frequency mentioned in the caveats section, but that is not the explanation we are looking for here, because of its small effect size.
- (3) In theory the length of green open access embargoes can also play an important role for the open access availability of publications. Embargoes are typically 6, 12, or 24 months, though some publishers use even longer embargo periods for their journals, or completely refrain from using embargoes (e.g. Emerald). However green open access data reported in Web of Science are restricted to green only, hence quite low. The effect on overall levels are thus expected to be minimal.

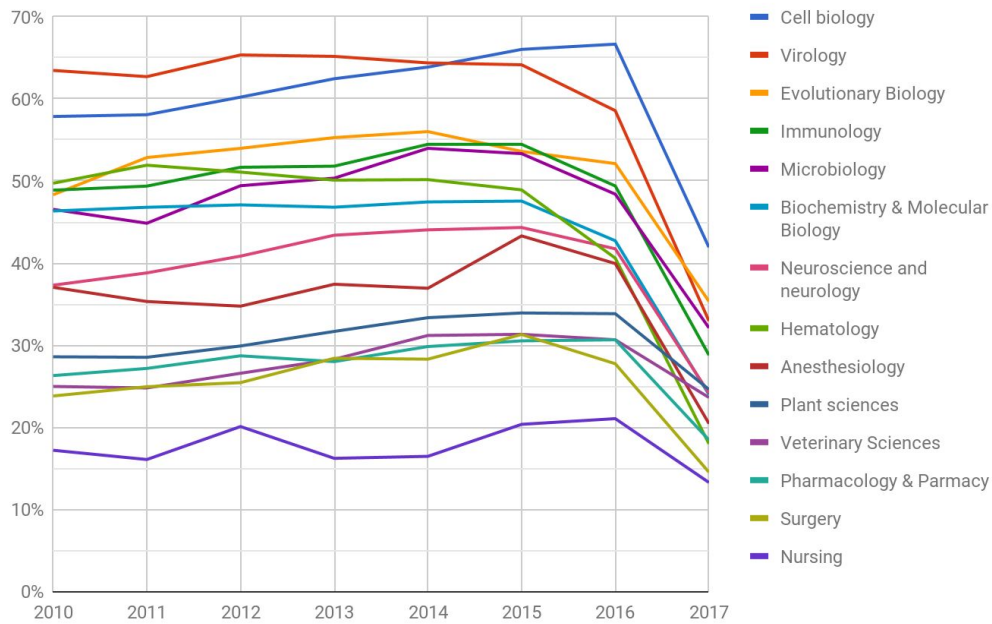
Fields

For fields we looked at the four main groups of disciplines. As there are over 100 fields in WoS it is impossible to display them all, so we made selections (see methodology section). Open access levels vary immensely by discipline, with the highest levels (over 50%) in some Life Sciences & Biomedicine and Physical Sciences & Technology fields and overall low (under 20%) levels in Social Sciences and Arts & Humanities (Fig. 3a-d).

Figure 3 Open Access levels of selected research fields 2010-2017 (articles + reviews) a. Physical Sciences & Technology, b. Life Sciences & Biomedicine, c. Social Sciences, d. Arts & Humanities.

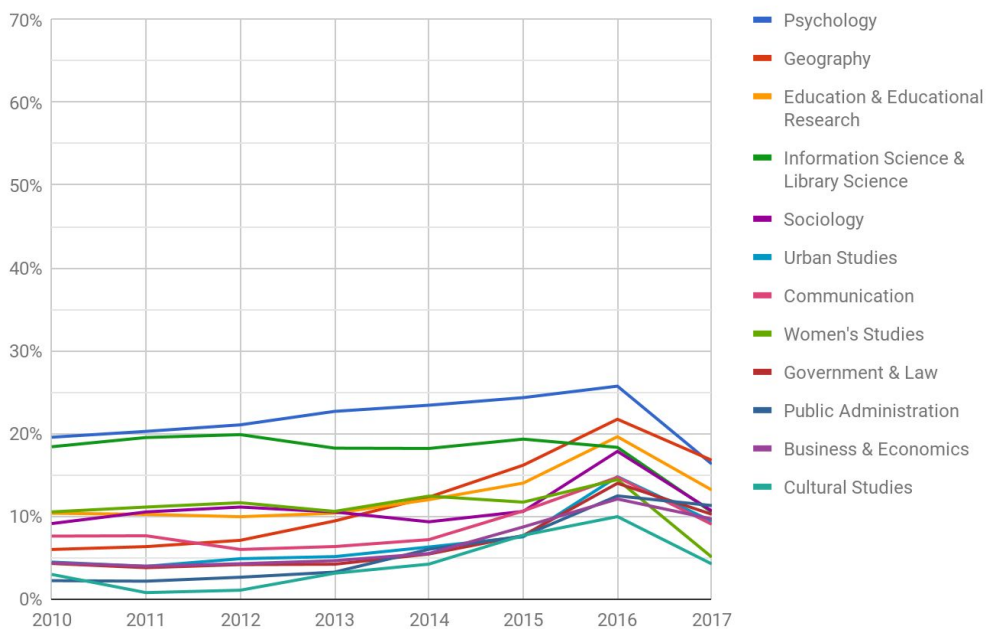


Life Sciences & Biomedicine



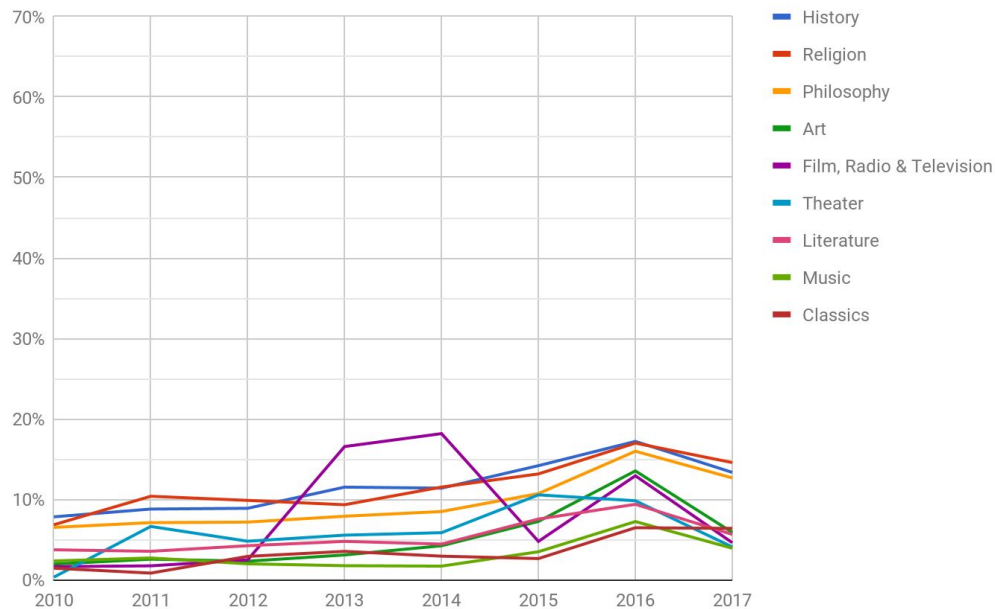
b

Social Sciences



c

Arts & Humanities



d

The range of levels *within* the broad disciplinary groupings is largest among Life Sciences & Biomedicine, while most Physical Sciences & Technology and Social Sciences are quite similar. Disregarding the 2017 anomaly, it is surprising to see more or less flat levels in Life Sciences & Biomedicine and Physical Sciences & Technology (with slightly lower levels for more recent years) but quite marked 'increasing' levels in Social Sciences and Arts & Humanities fields. There are some remarkable cases, such as the extreme differences between recent and older years for Astronomy & Astrophysics and between Physics on the one hand and Astronomy & Astrophysics on the other.

Generally speaking the potential explanations for different levels of OA could be:

- (1) Author behaviour: the degree to which authors choose to make their papers open either through green or gold open access routes. In turn that can depend on many things (acceptance of OA, availability of OA venues & infrastructure, availability of APC funding, APC levels etc.);
- (2) Publisher behaviour, with (a) the degree to which publishers decide to make full journal volumes or selected issues/papers open access and the timing of that and (b) the length of embargoes on green open access set by publishers.

There are clearly marked differences between fields regarding publication cultures and the status of available open access journals that may cause differences in author behaviour. Despite warnings against using impact factors to judge articles (DORA), this is still common practice and the lack of journals with (higher) impact factors in some fields may cause lower full gold adoption. In Arts & Humanities, less focussed on journal publishing anyway, there may be a lacking or lagging awareness of open access venues. Also in these fields the open access journals may be more spread out over numerous OA journals that are to a larger degree still outside of what Web of Science captures, compared to other fields where OA

publishing is perhaps more concentrated in a number of larger gold OA journals. The high levels we see for Psychology may be due to a publication culture more akin to Life Sciences & Biomedicine than to the other Social Sciences. The same holds for Library Science & Information Science, which is partly more akin to Computer Science in the Physical Sciences & Technology group of disciplines. For this field there is a remarkable lack of 'growth' of the levels of OA, with 2016 lower than 2010, something that we indeed see more often with the Physical Sciences & Technology fields.

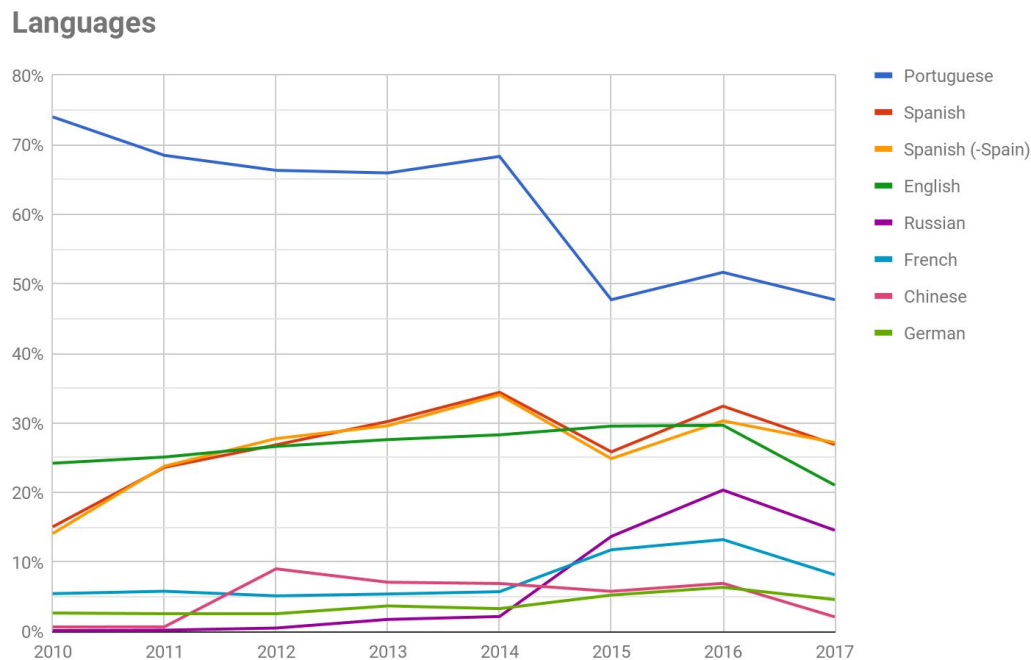
Some of the observed levels could also be related to publisher behaviour. One of the more common patterns in the data is the lower levels in recent years (again disregarding 2017). For instance for mathematics the sharp 'drop' in 2014 could be caused by the time lag of 4 years that Elsevier uses for making mathematics content open. Similarly part of the lower 2015-2015 levels for geochemistry and geophysics could be caused by the 24 month time lag that the American Geophysical Union uses for making its journals hosted by Wiley open. Another possible effect, this time in physics, is that of the [SCOAP³](#) deal that makes 10 journals open, involving ~5000 papers yearly. This deal could be responsible for higher levels of OA in physics from publication year 2014 onward. The difference between recent and older years in Astronomy & Astrophysics could also be an effect of diverging bronze OA policies. Because in these fields most authors already share their papers as preprint, open access is probably by many not seen as something necessary to invest in and left to publishers. Some of those (e.g. the American Astronomical Society) make journals open after 12 months, others only after many years or never.

Finally it is interesting to see in both physical and life sciences a range more or less from most pure to most applied sciences, with the former having the highest OA levels and the latter the lowest. Could it be that pure science funders are more strict? Is there more money in those fields to pay for APC's? Are publishers more reluctant to make the more applied journals open through the bronze route because of loss of income from non-academic subscribers? Without further research this remains guessing.

Languages

Although Web of Science records are 94.7% English (2016, down from 96.2% in 2014 largely because of the introduction of the Emerging Source Citation Index (ESCI)), it is still viable and interesting to look at OA levels for the larger source languages (Fig. 4). Of course the even relatively (compared to language composition of all scholarly output worldwide) dominant English language orientation of WoS coverage means that output capture by WoS will be more complete or more balanced for some countries than for others.

Figure 4 Open Access levels - selected languages 2010-2017 (articles + reviews).



Generally OA levels for English are fairly even for most years except 2017, and though more fluctuating, the levels for Spanish are comparable to those for English (with Spanish from Spain not behaving differently from Spanish from other countries). Portuguese (for over 85% attributable to output from Brazil) stands out with very high OA levels, especially up to 2014. The other languages show substantially lower OA levels.

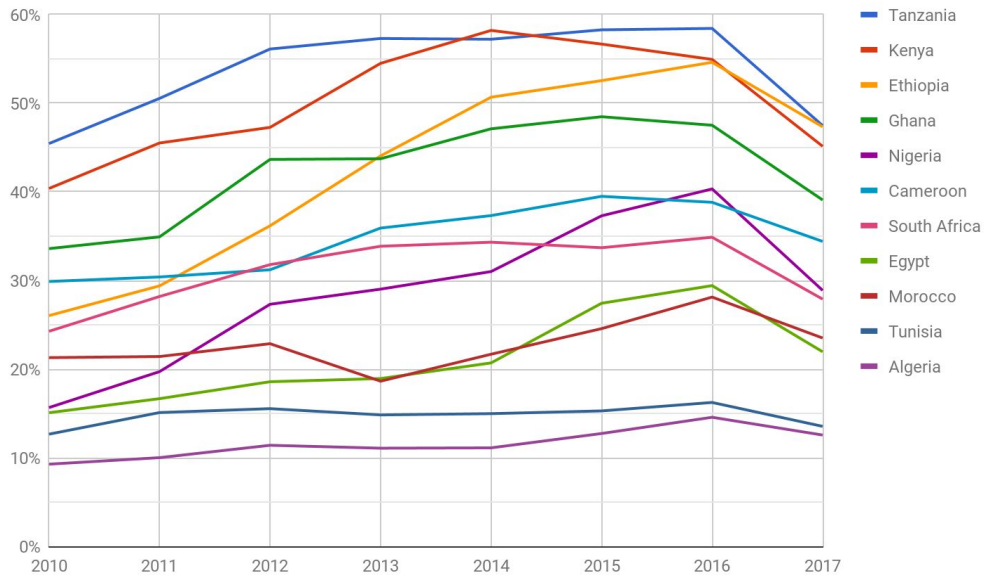
The picture for the languages other than English is more complex in terms of levels observed for individual publication years. The introduction of the ESCI likely contributed to changed levels for 2015 and 2016, but the pattern is not unambiguous, with 'increases' for French and especially Russian and 'decreases' for Portuguese and Spanish. It seems that languages that already had high OA levels in 2014 (Portuguese and Spanish) witnessed a detrimental effect on OA levels by the ESCI addition, while those with low levels were positively affected. The newly added ESCI has a larger percentage of OA journals (37%) than the full WoS Core Collection (16%, see Clarivate Analytics 2016a).

Countries

Open access penetration figures at country level are interesting because they can be related to country (and world-region) research policies and publishing traditions. They can also be compared to other claims about OA penetration in various part of the world (e.g. Kieñć 2017).

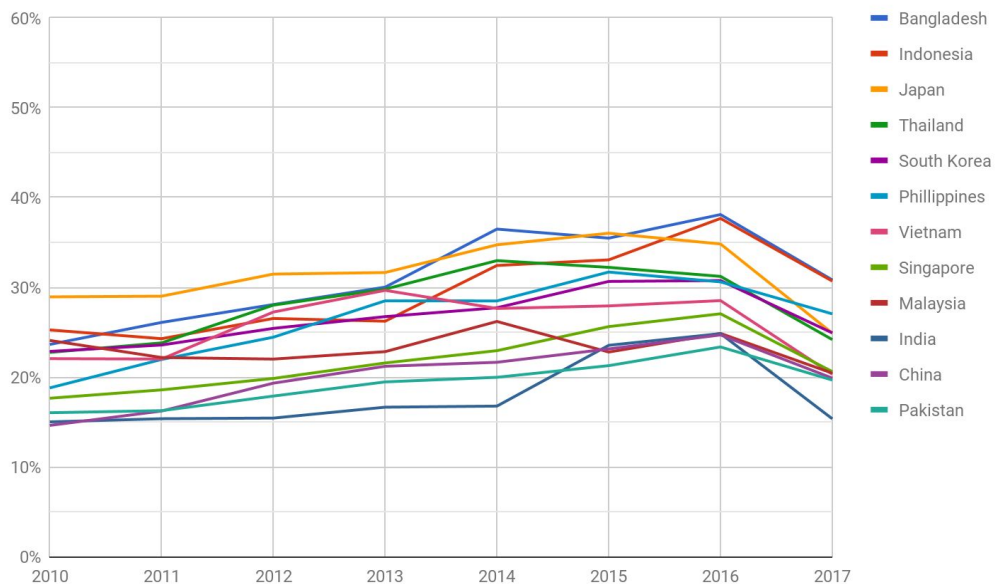
Figure 5 Open Access levels of selected countries 2010-2017 (articles + reviews) a. Africa, b. Asia, c. Europe (larger research output), d. Europe (smaller research output), e. Latin America, f. Middle East and Eastern Europe, g. Countries worldwide.

Africa



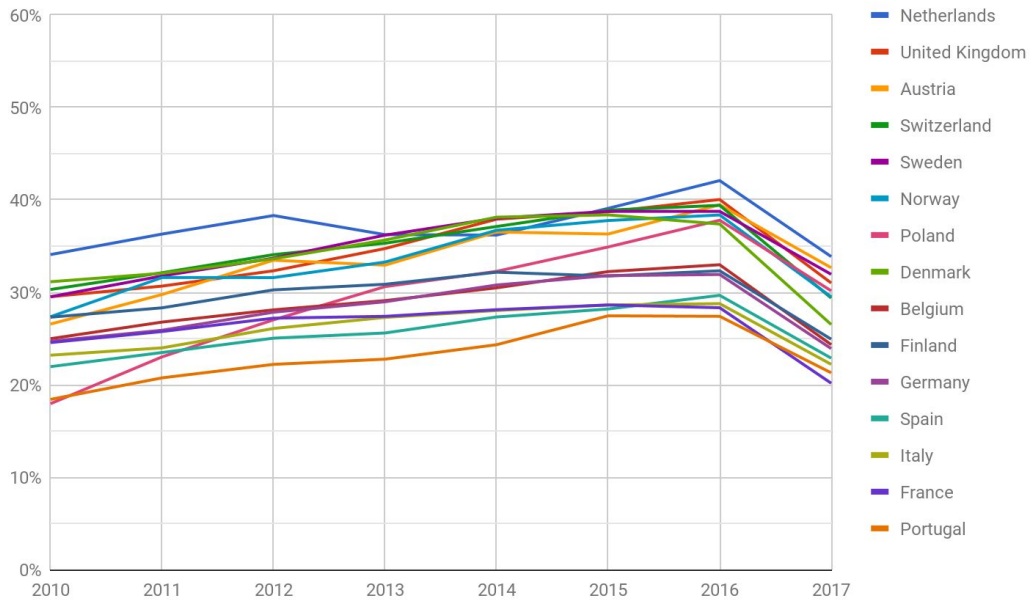
a

Asia



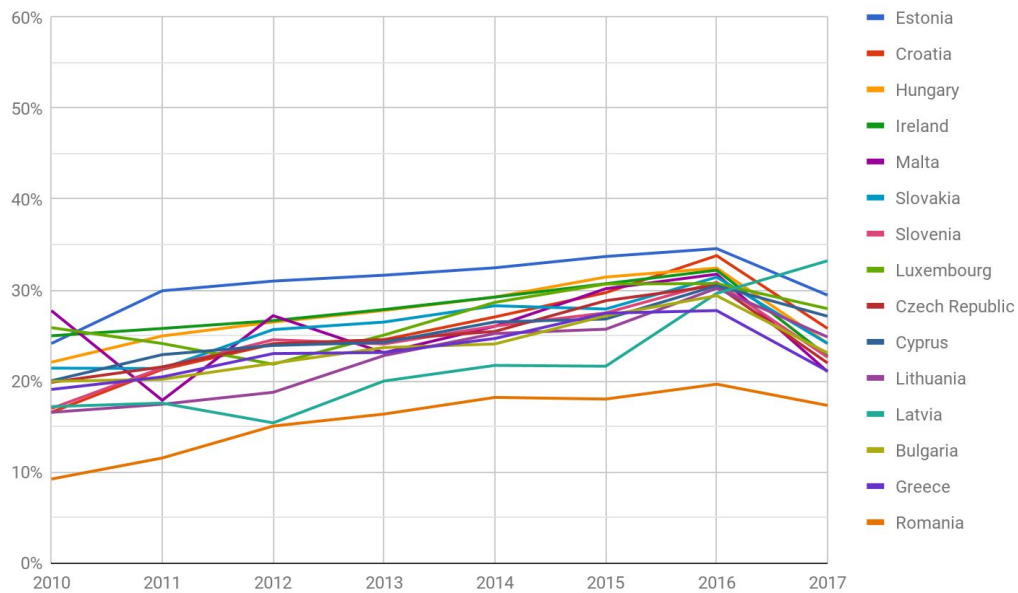
b

Europe (larger research output)



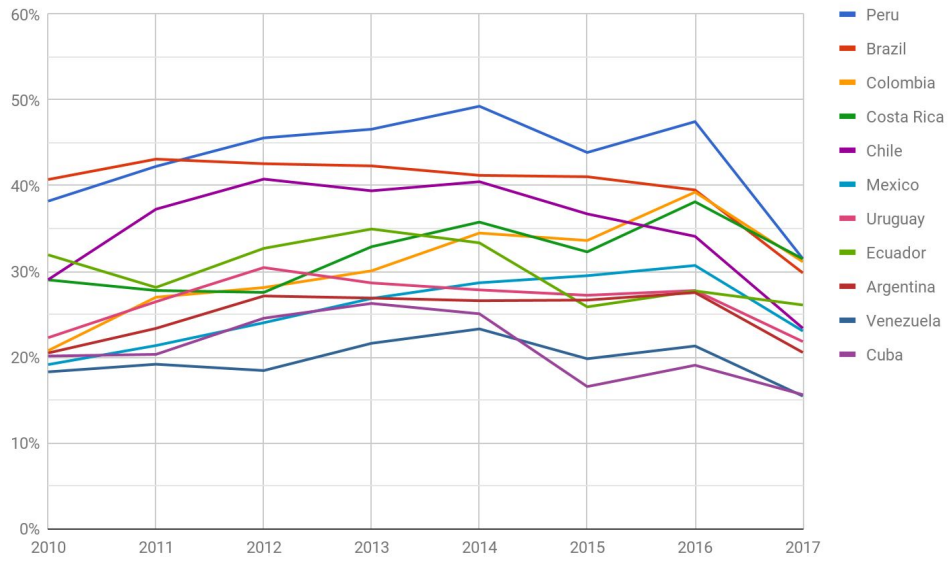
c

Europe (smaller research output)



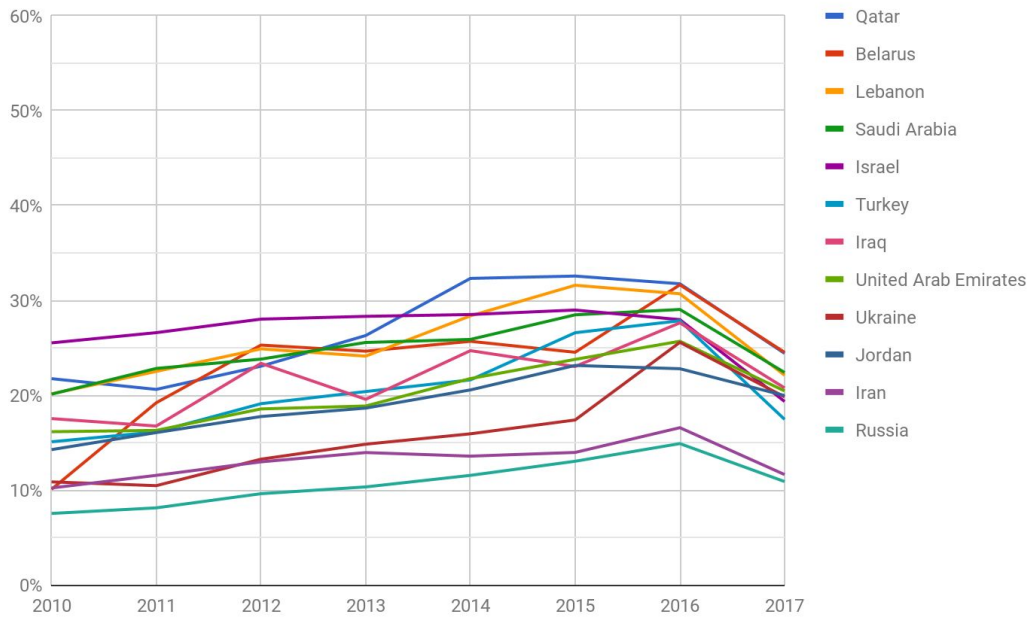
d

Latin America



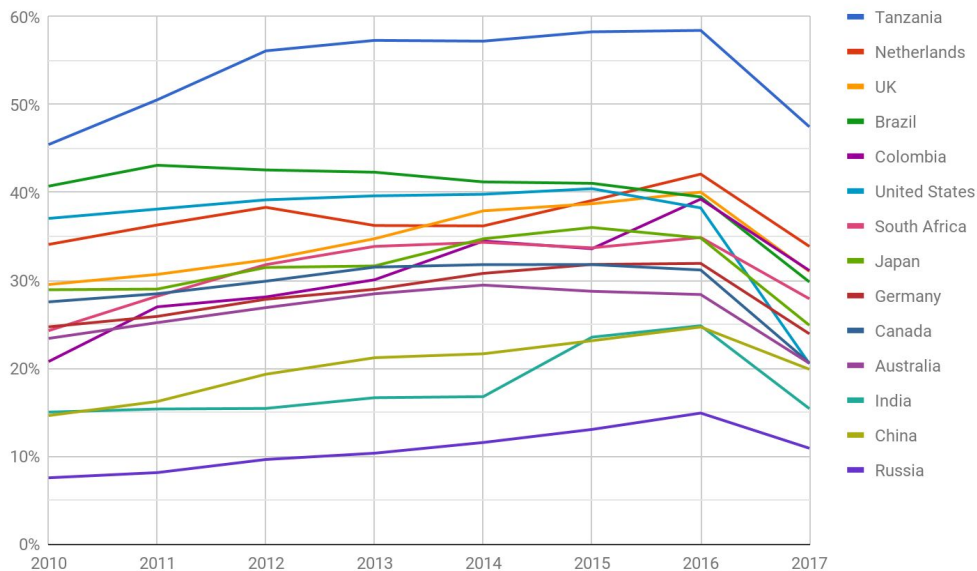
e

Middle East and Eastern Europe



f

Countries worldwide



g

In interpreting the data (Fig. 5a-g) we have to be aware of the varying degrees to which the research output of countries is captured by Web of Science. This is caused by Web of Science inclusion criteria that result in (some) bias against non-English, non-journal, open access, not highly cited, irregular appearing, nationally focussed publications. This is despite recent inclusion of the emerging sources citation index and despite steps to include more open access journals. Especially if the publications not covered can be expected to have open access levels diverging from output that is covered, the levels shown here may not fully reflect reality. A case that springs to mind is Latin America with many non-APC OA journals of which a large share is probably not covered by WoS (according to the Directory of Open Access Journals, 94% of the Latin American OA journals do not have an APC, compared to 70% globally). The composition by field of the output of countries can also influence the country levels: more Social Sciences and Arts & Humanities may result in a lower national open access levels. Also, countries that have many repositories not yet harvested by oaDOI may see their open access levels as reported here suffer. Finally, country data in WoS include research from all types of institutions: universities, government research institutes, but also commercial companies and labs.

Looking at all graphs (Fig. 5a-g), OA levels seem to be higher for each more recent publication year, except 2017, with 2016 some 10 percentage points higher than 2010. From this very rough general image there are many diversions. We find the highest levels, but also the widest range, in Africa and the lowest levels in Middle Eastern and Eastern European countries and to a lesser extent in Asian Countries. In Latin America en Europe the range is similar (~20% to ~40%). But of course these overall impressions do not do justice to the detail of the patterns for each macro-region.

In Africa (Fig. 5a), sub-Saharan countries (except South Africa) have the lowest research output but score the highest on OA, even higher than most rich countries outside Africa. Tanzania especially stands out. Arab countries in the North of Africa score much lower. We could speculate about a possible language effect (Arabic & French versus English), but without further research it is difficult to say what is going on here. The only thing we know is that the explanation is not in high levels of green only OA. South Africa has according to WoS the largest output volume of all African countries, only approached by Egypt, but OA levels for the recent publication years are flat. It is interesting to note that many of these African countries already had high OA levels in 2010-2014, so before the addition of ESCI.

The Asian pattern (Fig. 5b) also shows there no clear-cut relation between wealth and country OA levels. Yes, Japan is high here, but is surpassed by Indonesia and Bangladesh for 2016, and Singapore is among the lower half. Some of the more abrupt changes in 2015 are possibly due to ESCI inclusion. It depends on the OA status of specific journals from the region that are included in ESCI whether the effect is positive (as perhaps for India, Philippines, South Korea) or negative (as perhaps for Malaysia, Thailand, Bangladesh).

European countries have overall 2016 levels ranging from 27-42% in 2016 (excepting Romania (Fig. 5c-d)). That is substantially lower than the levels from 1Science that the European Union works with (ScienceMetrix 2016): 52% overall, 21% gold and 45% green only OA in 2014. It is interesting to see that in Europe most of the the smaller countries (measured by research output in WoS) lag some 5-10 percent behind the larger countries. This distinction is also crudely the distinction between Southern and Middle European countries and Western and Northern European countries (with notable exceptions). This may reflect greater reliance on yet incompletely harvested output shared via green routes in Southern and Middle Europe (see Johnson 2017, p.44). But it could also be partly attributable to lower ability to pay APCs and/or to weaker open access requirement from funders. For the larger countries, looking at the 2016 levels there seem to be two groups, separated by some 7 percentage points. Interestingly Poland is in the lower group for the earlier years but 'joins' the higher group for the recent years. Another interesting case is Latvia with a very steep 'rise' for each more recent publication year. Latvia is one of the very few countries that manages to have higher levels for 2017 compared to 2016. This is even more surprising as Latvia has no open access policy (Ministry of Education and Science of Latvia 2016). Finally, the UK data are quite nicely in line with what was recently reported for the UK (Research Information 2017), where immediate OA levels of 25% were found, rising to 33% after 24 months (though that was on a sample of universities' output only).

Despite probable underreporting it is clear that the EU countries will not manage the 100% open access in 2020 they agreed to in 2016 by merely relying on the 'trend' (10% increase between publication years 2010 and 2016, i.e. 1.5 percentage point per year). Even if our underreporting was 50% (so levels would be on average 45% instead of 30% in 2016, the year-on-year 'growth' of OA levels rate to reach 100% would need to increase from 1.5 percentage point to almost 15 percentage points. The need for increased OA growth rates to reach the goals has also been signalled by others as well (e.g. Johnson 2017). To reach this a real break in policies, infrastructure and compliance would be necessary. Of course the first year for the policies and goals set out in 2016 to produce results would be 2017, so we

will have to wait until early 2019 to be able to fully (accounting for articles made available after 6-12 month embargo periods) evaluate that with these WoS/oaDOI data. Also for the effect of other recent mandates and policies, like the English HEFCE open access mandate voor REF-eligible articles for output accepted from April 1 2016, we will have to be patient to see the results.

Latin America is well known for its 'third' way in open access: non-paywalled, non-APC based but open journals funded by institutions, national funders and government bodies. Indeed some countries (Brazil, Chile and especially Peru) show high OA levels (Fig. 5e), though probably still below real levels due to possibly unharvested repositories. But Latin America according to these data is also a mixed bag: some countries show lower levels for each consecutive publication year after 2012 (Argentina, Uruguay, Chile, Brazil). Perhaps it's the drive to publish in more English but without the funds to pay expensive APCs? The observed 'declines' in 2015 may have to do with the introduction of ESCI in various ways. ESCI introduces Latin American journals publishing in Spanish and Portuguese but a simple count in WoS indicates that papers in those journals are less often open access in new journals added by ESCI than Spanish and Portuguese papers in journals that were already in WoS as part of the other constituting databases (SCI, SSCI and A&HCI). That seems to be in line with what Carvalho Neto (2016) found for Brazilian journals: higher rated journals (that may already have been in WoS before 2015) are more often full gold than middle and lower rated journals (that may have been added to WoS through ESCI).

The Middle Eastern and Eastern European countries (grouped together for practical reasons rather than being a coherent region) have on the whole the lowest levels of all macro-regions (Fig. 5f). Not unexpected are the low levels of Iran and Russia. Some other countries, like Belarus, Qatar and Lebanon show remarkably high levels for recent publications years compared to the older years. The complexity of drivers of OA levels are also apparent from that, as the backgrounds of these three countries in terms of wealth, languages and likely also in academic cultures and research policies are so different, while they are displaying very similar OA levels in 2016.

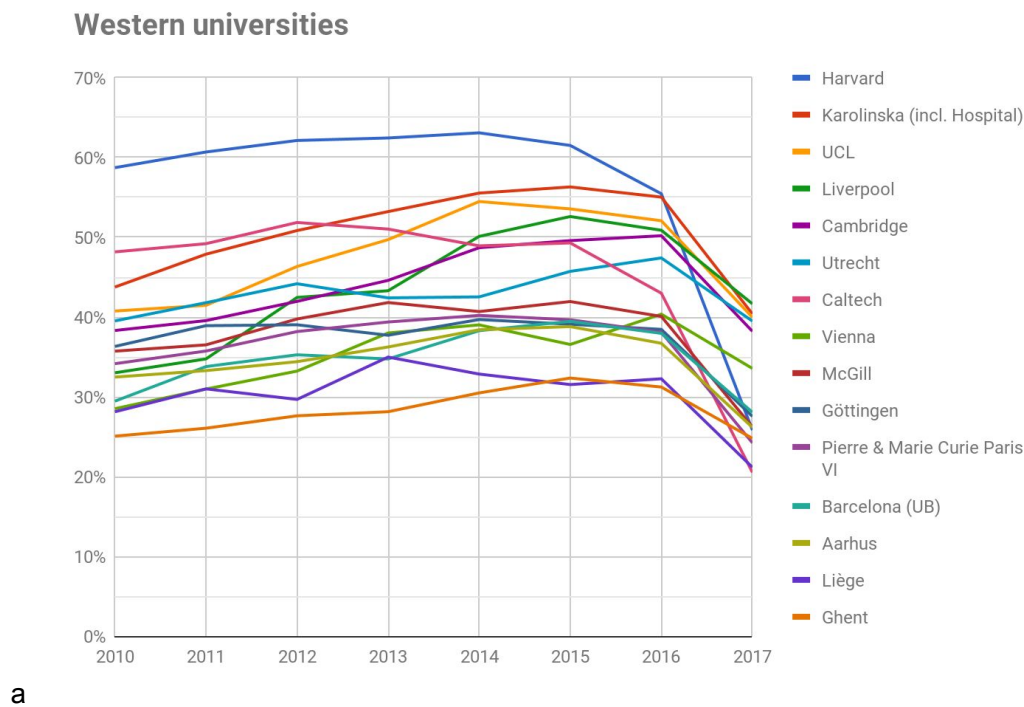
Countries from North America and Oceania are included in a final worldwide graph (Fig. 5g). Strikingly, Canadian and Australian levels are somewhat mediocre in international perspective. US levels are high, not unexpectedly with its strong funder (e.g. NIH, its largest biomedical research funder) and institutional open access mandates.

The global comparison learns that US and Dutch levels, though high, are still lower than Brazilian and Tanzanian levels for most publications years. In line with the country's reputation as vocal OA proponent, the Dutch levels are indeed among the highest of rich Western countries according to our current data. We will explore the Dutch situation further down.

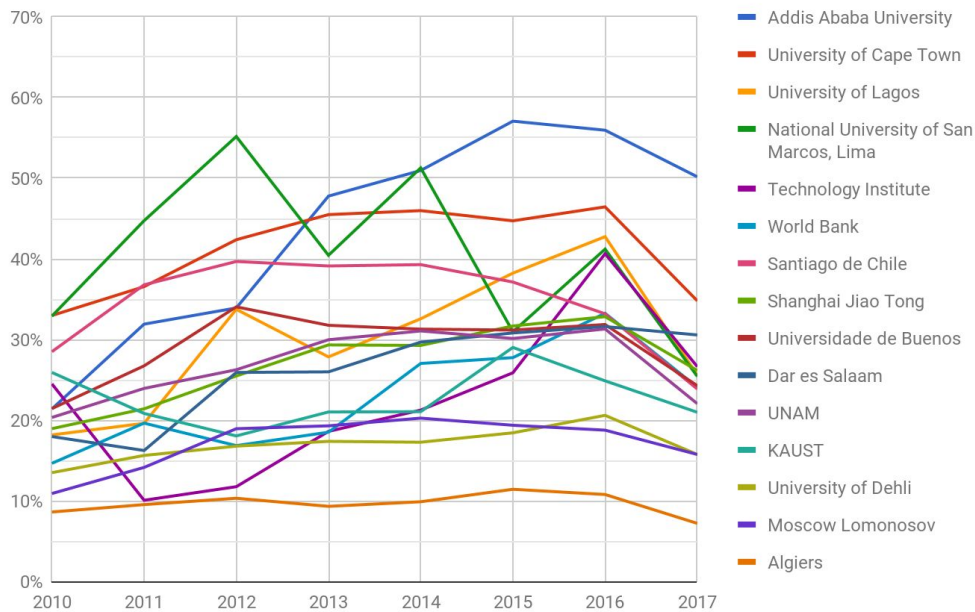
Institutions

There is big diversity in the height and development of the open access levels of (mostly rich) Western universities and non-Western universities and institutions alike (Fig. 6a,b). Again, we have to be careful in drawing fast conclusions, because the data have not been put to the test in this way before, and especially here the absence of harvesting of some institutions' repositories can have great repercussions for the levels found.

Figure 6 Open Access levels of selected institutions 2010-2017 (articles + reviews) a. Western universities, b. non-Western universities & World Bank.



non-Western universities & World Bank



b

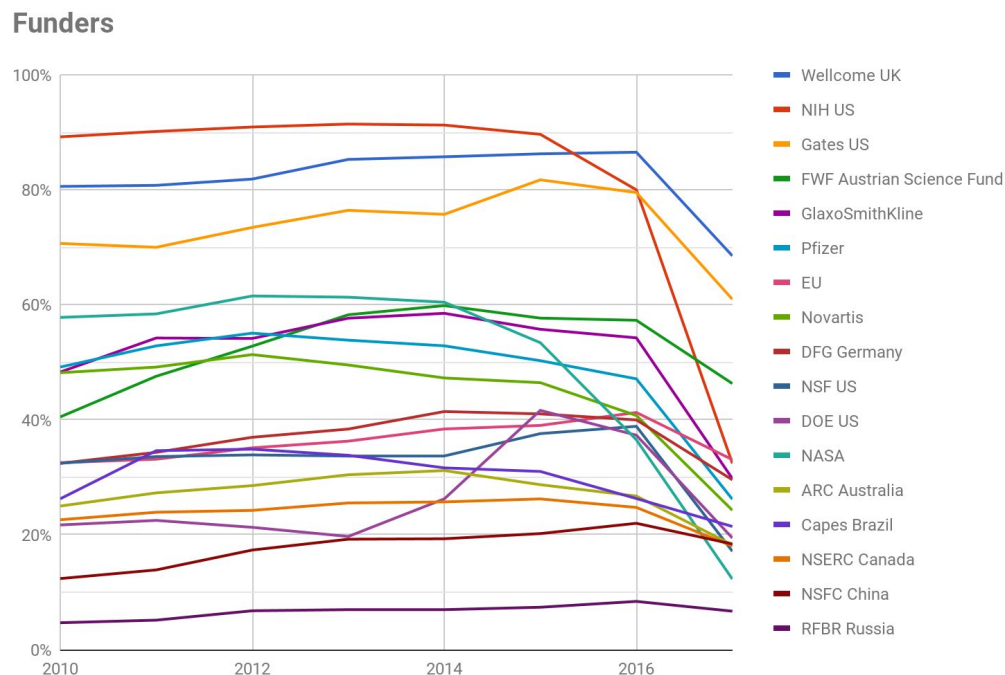
Looking at the selection of Western institutions (Fig. 6a) some of the usual suspects like Harvard and Cambridge are found at or near the top. The low level for Liège, famous for its strong green open access mandate, its ORBi repository and a deposit compliance rate of 82% (Université de Liège 2017) raises some doubt over the data provided here by WoS/oaDOI. According to their own statistics the articles *added* to the Liège repository ORBi were 59% OA in 2016 and 53% in 2017 (ULiège Library Network 2018), compared to 32% in our WoS data. Could the 19 percentage point difference be explained only by the fact that Liège publications indexed by WoS are less frequently open access compared to all publications in their repository?

Most institutions from the global south (to which we add a few others for comparison, see Fig. 6b) show levels that reflect country averages: high levels for institutions in sub-Saharan Africa, with for instance very high levels in Addis Ababa, but lower levels in Asian and the Middle Eastern institutions. Two interesting cases are Dar es Salaam, with substantially lower levels than Tanzania as a whole, and the Technology Institute in Bandung that shows markedly higher levels for each more recent publication year. The curve shapes are much more volatile than for the western countries, this may be caused by sometimes relatively low absolute numbers for the earlier years (<50, especially for Algiers, Dar es Salaam and Bandung) and by the addition of ESCI affecting publication years from 2015. Finally, it is somewhat shocking that 3 out of 4 World Bank authored articles are not open access according to these data.

Funders

For this first simple exploration we just looked at the 16 of the 100 most often mentioned funder names and added the Gates Foundation and the Austrian Science Fund for comparison (Fig. 7).

Figure 7 Open Access levels - selected funders (articles + reviews).



The smallest funder (Gates Foundation) in the analysis had just over 2400 funded papers in 2016, the largest (NSFC from China) had over 194 thousand (more than all the others combined). These absolute figures are an underestimation not only because not all records have funding information but also because we only grouped name variants of the first 100 mentioned funder names.

The most prominent feature of the various OA levels for funders is on the one hand their enormous range and on the other hand their stability across years of publication. That large funders from Russia and China show low levels is not unexpected, but we also find large funders from Canada and Australia in these lower bounds. Wellcome (UK) and NIH (US) show high levels, not surprising given their mandates. The longstanding open access mandates of NIH and Wellcome, requiring deposit in PubMed Central or Europe PMC (that are fully harvested by oaDOI for detection of green only OA) within 12 and 6 months after publication respectively for all research funded, even begs the question why that level is not 100%. The drop in 2017 is of course partly explained by those 6 and 12 months. But for the other years it is unclear why levels are below maximal. Is the policy not fully complied with or is some OA not captured by oaDOI? For Wellcome (not for NIH) part of the explanation may

lie in review articles, that are not subject to the Wellcome OA requirement. In the data we used review articles make up about 10% of the Wellcome funded output. But many of those are published open access as well. Excluding reviews would make the level of Wellcome rise only about 2 percentage points.

Interestingly there are high levels for some pharmaceutical company funders as well. The EU is in the middle ranges. The odd ones out are NASA and DOE, the US department of energy. The first has strikingly low levels for recent years and the second in contrast has clearly heightened levels from publication year 2014. A final interesting case is that of the Austrian Science Fund (FWF) that has open access levels around 60%. This is remarkably lower than what they report themselves (83% and 92% for 2015 and 2016 respectively, see Kunzmann & Reckling 2017). The background might be output not covered by WoS or missed by our search but with very high OA levels and perhaps also lack of green only detection by oaDOI. Because the FWF has shared their data and methods, this is a good opportunity to mutually corroborate the data.

All these levels should be interpreted with caution, because some funders may explicitly allow or even promote green open access and oaDOI might still be incomplete in harvesting institutional repositories in which many green versions reside. Further research with updated oaDOI information is needed to verify these levels and confirm the image of lack of 'growth' between 2014 and 2016.

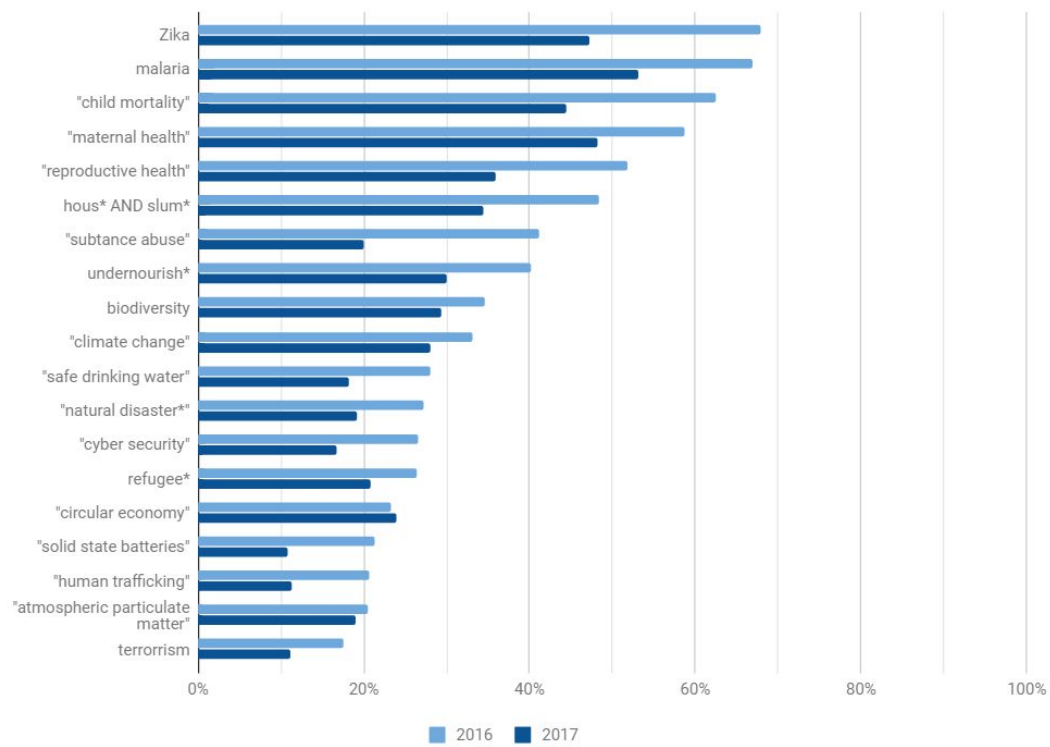
Topics

Open access aims at speeding up and improving research by making results to build on more widely available. In the end it is about making research results available to contribute to solving societal and technical challenges. That's why it is interesting to also look at open access levels of the body of literature pertaining to urgent (global) issues. We subjectively selected some topics, in part inspired by the UN millenium goals, and searched those. OA levels for these topics are depicted in Fig. 8.

Some of these subjects are clearly interdisciplinary/multidisciplinary, so OA levels are subject to OA-practices in various fields. Another caveat concerns the lower 2017 levels. Insofar as these are explained by embargoes on green or bronze OA, that points at an important drawback of those embargoes. However, if there are technical reasons for those lower levels those should be taken into account when interpreting the results.

Figure 8 Open Access levels of selected topics 2016-2017 (articles + reviews).

Topics (including UN millennium goals)



In line with OA levels found by field, levels above 50% are rare, meaning that for most of the research results needed to solve pressing issues, users need to pay or resort to sites as ResearchGate or SciHub. Worst off are the non-health issues. Almost nine out of ten recent papers on topics like solid state batteries (one of the candidates to make flying using renewable energy a reality) and human trafficking are not openly available. Levels are not much higher for most other topics badly in need for effective policies are technical solutions.

Even the disastrous Zika epidemic, despite wide attention, could not force publishers and the research community to make more than half of the 2017 papers on the disease openly available. There has been considerable attention for the problem, also around sharing data and coordinating research, including action to create a Zika corpus as Wikidata project (Wikidata 2017), but apparently still not enough.

A detailed look at open access levels: The Dutch case

Our comparisons so far looked at overall open access levels as available from Web of Science, making no distinction between green only and gold open access. In the data presented so far, also no distinctions are made between articles in full open access journals (pure gold OA) and hybrid journals (hybrid OA), and between journals publishing open access articles and those making articles free to read, often after an embargo and with no re-use rights (bronze OA). Since open access policies of governments, institutions and funders are often geared towards stimulating one or more types of open access specifically, the ability to break down open access levels into different types of open access will help assessing the effect of such policies. It will also give more information about developments in open access in general across research fields, institutions, countries, languages, funders etc. Here, we present the results such an analysis for a subset of our data, supplementing information available oaDOI in Web of Science (levels of all gold and green only open access) with information from the oaDOI API itself (levels of pure gold, hybrid and bronze open access).

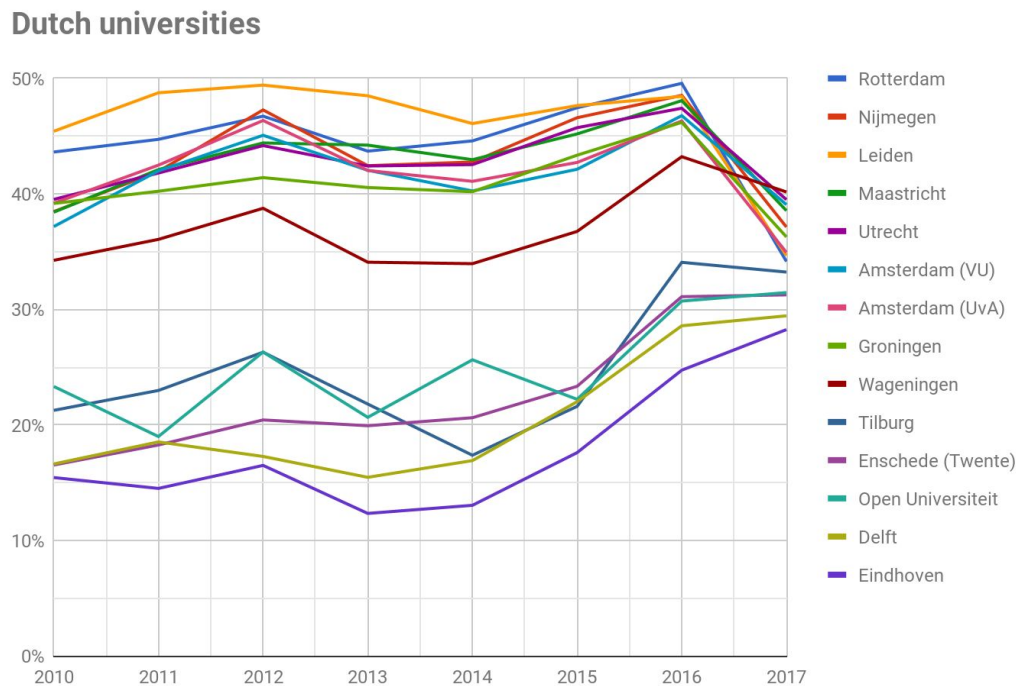
We chose open access levels at Dutch universities as the subject of this analysis, as it is a reasonably sized dataset to handle and offers the opportunity to assess the effect of Dutch national and institutional open access policies, including a clear goal set by the national government, a (non-binding) OA mandate by national research funder NWO and a large number of collectively negotiated big deals with publishers that include APC offsetting clauses for hybrid open access publishing.

In November 2013, the Dutch government announced the goal to strive for 60% open access publishing within 5 years, and 100% within 10 years, with a preference for gold and hybrid open access (OCW 2013). In 2017, this ambition was raised to achieving 100% open access in 2020, as formulated in the National Plan Open Science (OCW 2017) that was signed by, among others, the Ministry of Education, Culture and Science (OCW), the Royal Netherlands Academy of Arts and Sciences (KNAW), the Association of Universities in the Netherlands (VSNU) and the main public funders (NWO, ZonMW).

Open access levels at Dutch universities

For Dutch universities, we see a clear distinction between general universities (Rotterdam, Nijmegen, Leiden, Maastricht, Utrecht, Amsterdam (UvA and VU) and Groningen), that cover all disciplines, and the technical universities in Enschede (Twente), Delft and Eindhoven (Fig. 9).

Figure 9 Open Access levels of Dutch universities 2010-2017 (articles + reviews).



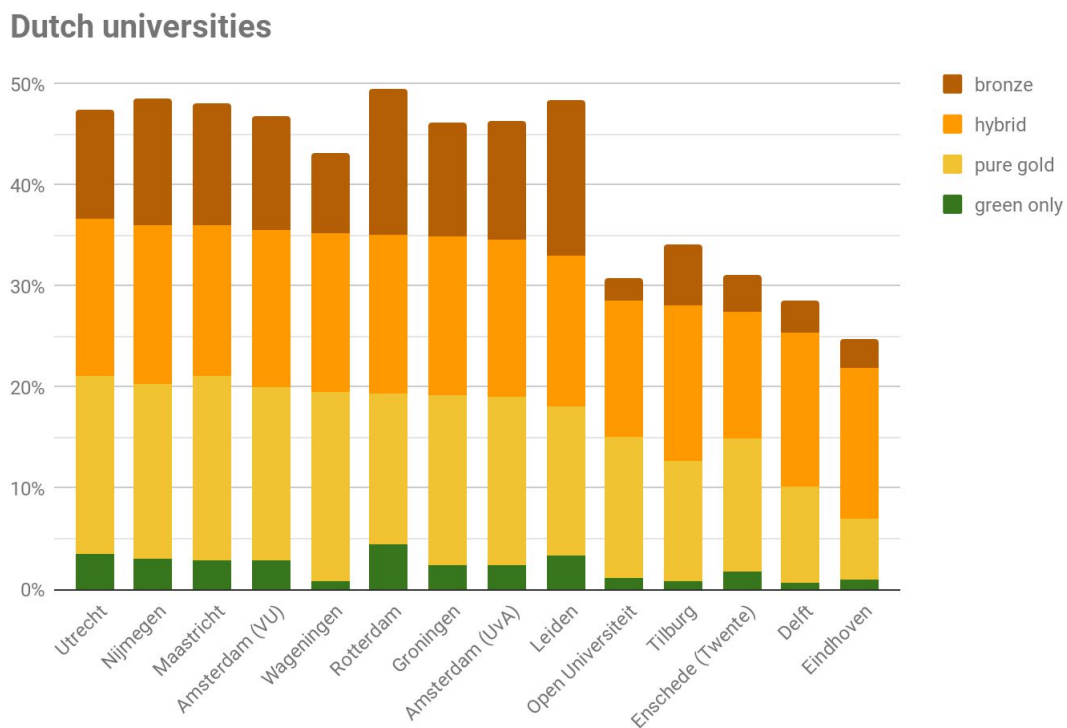
The general universities have higher overall open access levels, likely to do with the coverage of Life Sciences & Biomedicine fields at these institutions. This would also explain the lower levels of open access found for the two general universities that do not cover these disciplines: University of Tilburg (no Life Sciences and Medicine, and also no Physical Sciences) and the Open University (no Life Sciences & Biomedicine). Wageningen University, with a strong Life Sciences & Agriculture focus but lacking a medical faculty, occupies a middle position.

An interesting observation is the drop in open access levels in 2013 and 2014. This characteristic, consistent across almost all Dutch universities, is not found in any of the other parameters studied, including other rich Western countries and institutions (Fig. 6a). All Dutch universities show a consistent increase in open access levels in 2015-2016, with the increase being especially steep in the group of technical universities. This group also shows a less steep 2017 drop in open access levels. The volatile pattern seen for Open University (though overall consistent with the general trends described above) can be attributed to the low number of publications from this institution.

To extend this general analysis of open access levels, we analyzed levels of green only, pure gold, hybrid and bronze open access for all 14 universities in the Netherlands for publication year 2016, combining information from Web of Science with information from the oaDOI API (Fig. 10). This allows us to assess the contribution of various types of open

access and the differences therein between universities, and is also potentially relevant for assessing national and institutional open access policies.

Figure 10 Open Access levels (green, pure gold, hybrid and bronze) of Dutch universities 2016 (articles + reviews). Universities are ordered by total level of green+pure gold+hybrid OA.



The data show relatively high levels of bronze open access for general universities compared to technical universities. Consequently, the difference in open access levels between these two groups of universities is much smaller when only green, pure gold and hybrid open access levels are taken into account. Among general universities, two (Leiden and Rotterdam) have an especially large percentage of bronze open access. Levels of green only open access are generally low, but somewhat higher for universities with a medical faculty can also be observed across this full Dutch sample.

The overall percentage open access for peer-reviewed articles from 2016 for all Dutch universities as reported by the Association of Universities in the Netherlands (VSNU 2017) is consistent with the overall figure for the Netherlands in Web of Science data (see Fig. 5c,g) (both 42%). Note that the figure for the Netherlands as a whole also includes output from other institutions, like government research organizations, companies and commercial labs. At the level of individual universities, it will be interesting to compare the results of this unified approach using Web of Science data with the numbers collected by the universities. This would also be an interesting test case to compare the detection of green OA from institutional repositories by oaDOI with numbers reported by universities themselves.

Breakdown of longitudinal open access levels for selected universities

In an attempt to explain the observed patterns in open access levels, we selected four Dutch universities (two general, one technical, and one mixed type), and investigated the levels of green, pure gold, hybrid and bronze open access for the years 2010-2017 (Fig. 11a-d, Fig. 12a-d). We chose Leiden University and Utrecht University as examples of a general universities, Technical University Delft as an example of a technical university, and included Wageningen University as a mixed type.

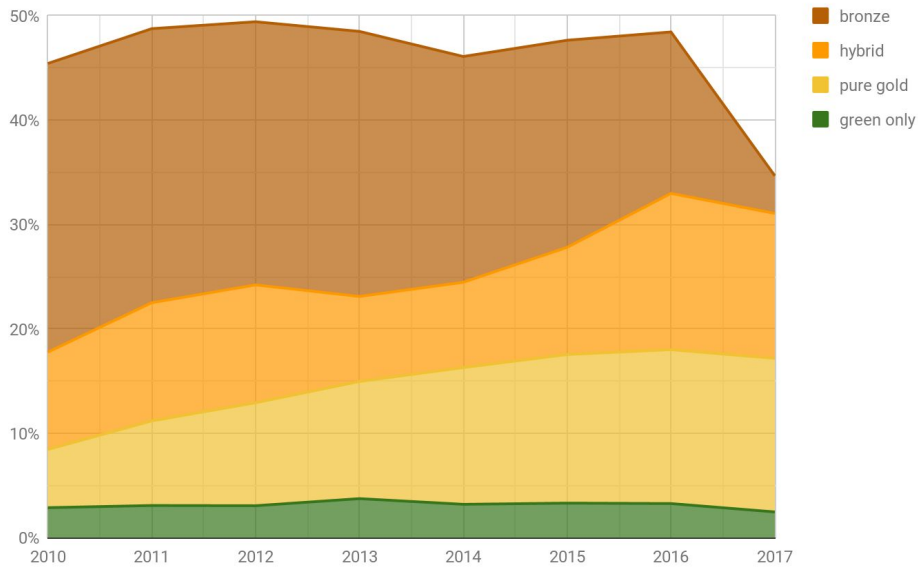
Pure gold and hybrid open access

For all universities investigated, there is a steady but gradual increase in the level of pure gold open access for 2010-2016. The observed dip in open access levels in 2013-2014 can be attributed to hybrid open access levels. Hybrid open access levels rise relatively steeply in 2015-2016, compared to levels of gold open access. In our sample, the increase in hybrid open access in 2015-2016 is most steep for the technical university, where hybrid open access had overtaken pure gold open access in 2016.

This rapid increase in hybrid open access publishing is consistent with the timing of the collectively negotiated *big deals* in the Netherlands that include a provision for open access publishing in hybrid journals, with no APCs payable by corresponding authors from Dutch universities and academic medical centers. In 2015 and 2016, such *big deal* agreements were made with the larger publishers (SpringerNature, Elsevier, Taylor & Francis, Wiley, SAGE (Openaccess.nl 2017)). The rise in hybrid open access publishing in 2015-2016 closely follows this timeline. An explanation for the observed halt in growth of hybrid OA in 2017 could be that in this year, agreements were added, but only with smaller publishers. Thus, the big increase in hybrid OA publishing in 2015-2016 is not expected to be repeated in 2017.

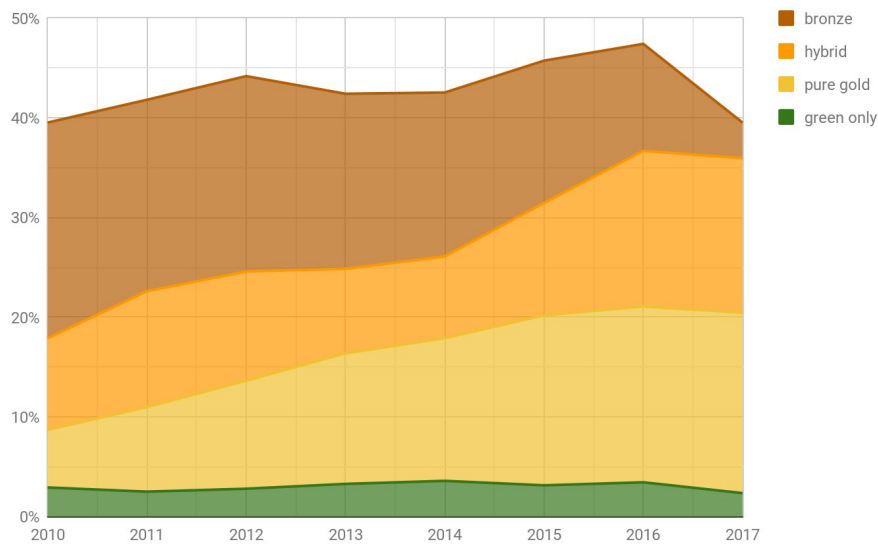
Figure 11 Open Access levels (green only, pure gold, hybrid and bronze; stacked) of selected Dutch universities 2010-2017 (articles + reviews). a. Leiden, b. Utrecht, c. Wageningen, d. Delft.

Leiden University



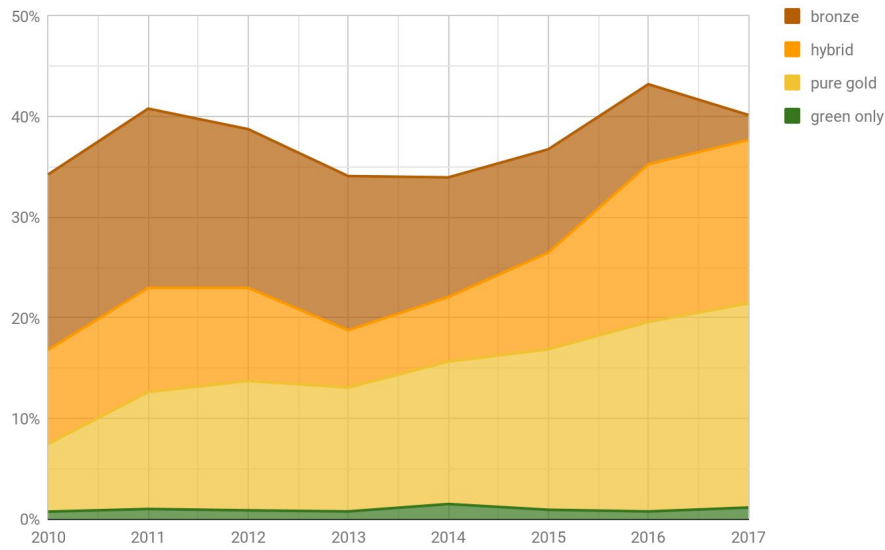
a

Utrecht University



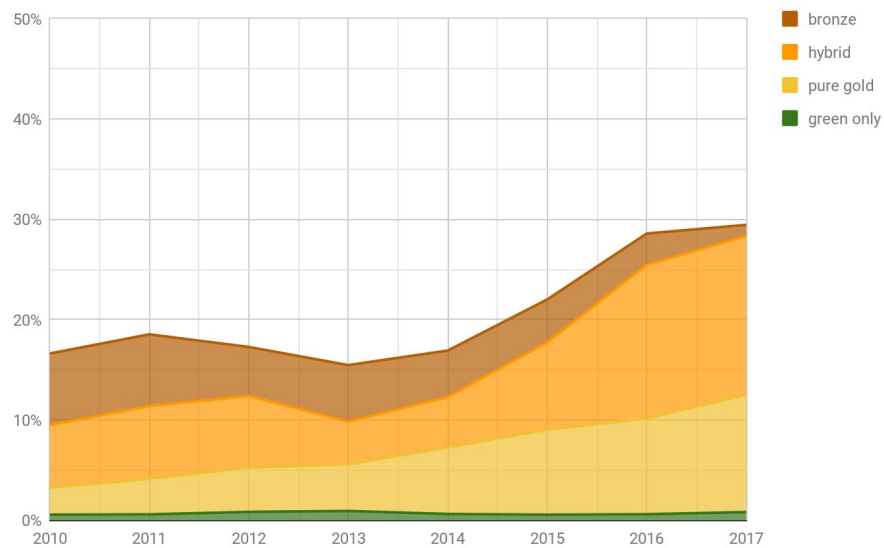
b

Wageningen University & Research



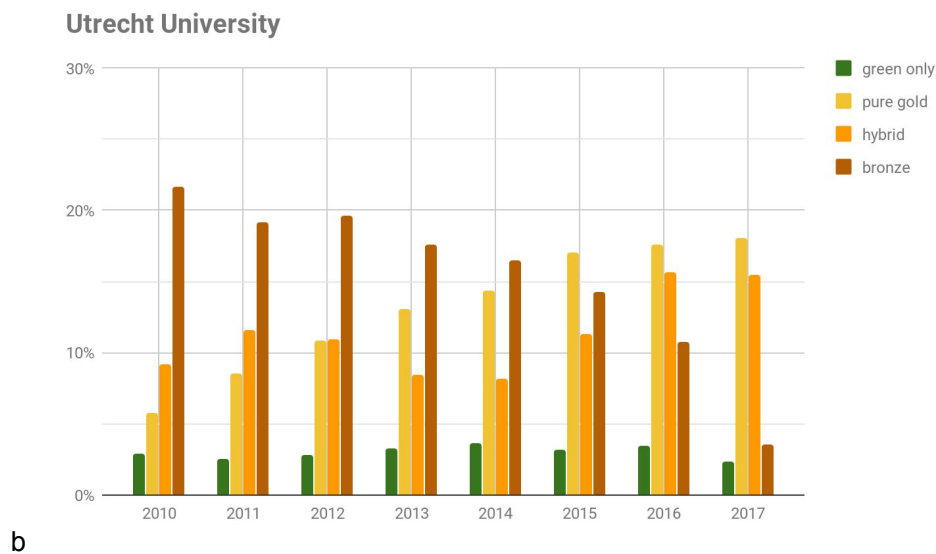
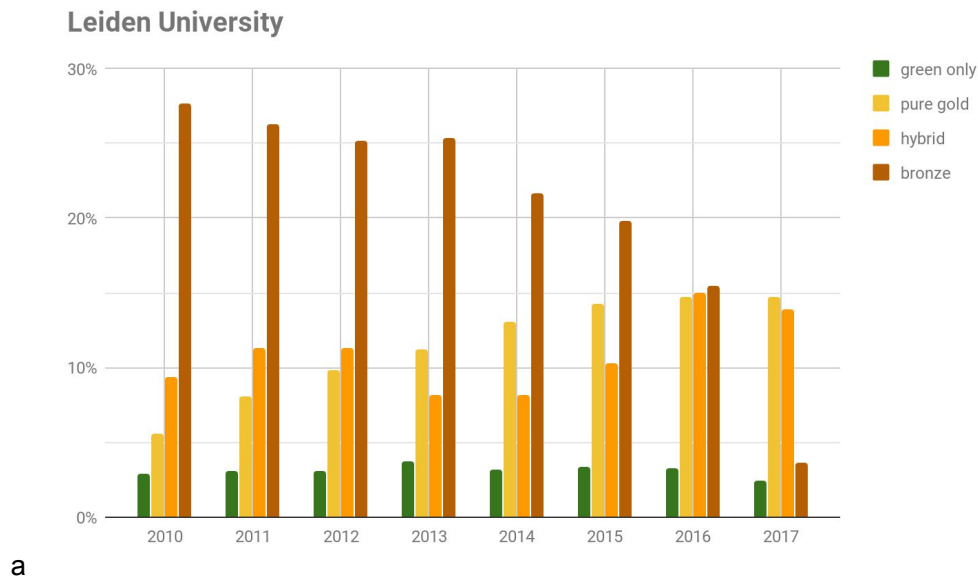
c

TU Delft

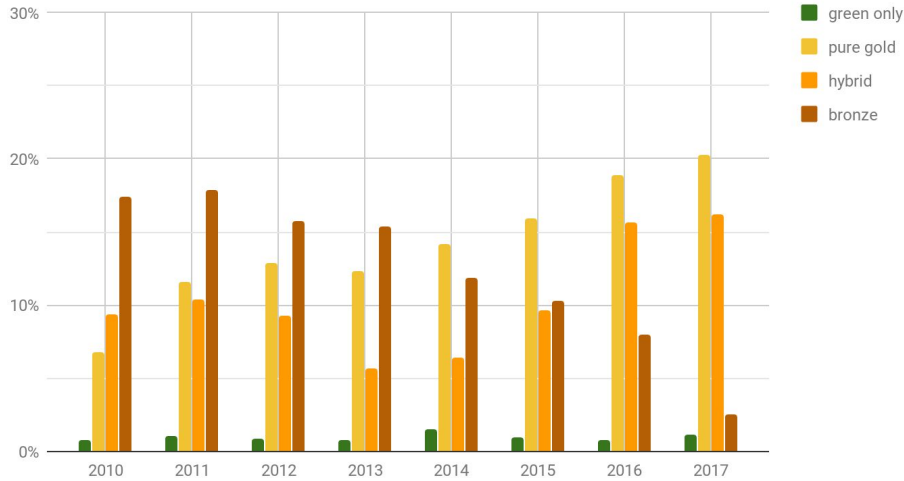


d

Figure 12 Open Access levels (green only, pure gold, hybrid and bronze) of selected Dutch universities 2010-2017 (articles + reviews). a. Leiden, b. Utrecht, c. Wageningen, d. Delft.

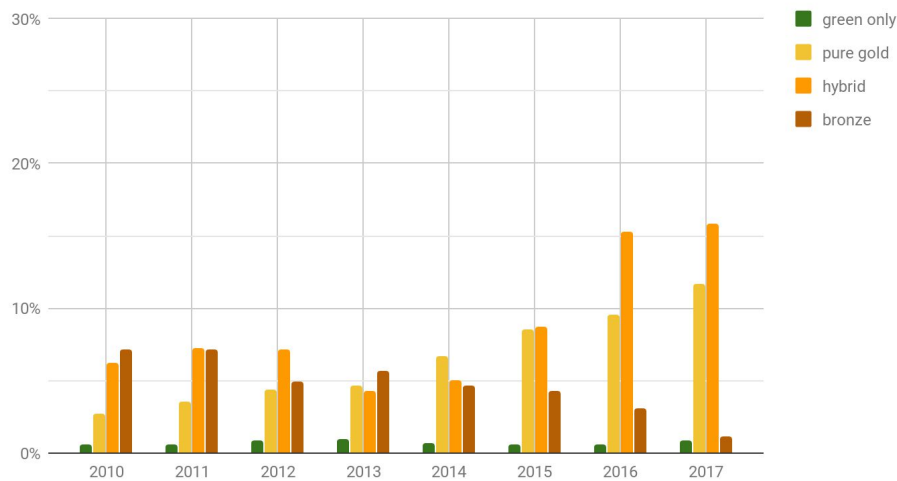


Wageningen University & Research



c

TU Delft



d

It should be noted that not all hybrid open access publishing in the Netherlands is covered under these *big deals*. First of all, not all publishers are yet participating, and the share of journals from a publishers' portfolio that is included in these APC-offsetting agreements varies considerably. Second, not all authors are yet aware these agreements exist, so some will likely have published open access in these journals without realizing they would not have had to pay the APCs (and just as likely, many will have refrained from making their articles in these journals open access for this reason). Finally, the agreements only hold for corresponding authors from Dutch institutions, so hybrid open access articles on which authors from Dutch universities are co-authors, but not corresponding author, will not have been part of these deals.

Possibly, a rising cost of hybrid open access publishing prior to these deals coming into effect is to blame for dip in hybrid open access levels in 2013-2014, but this needs to be corroborated [by comparing the observed trend with \(hybrid\) APC-levels over time](#). Moreover, it would not explain why for all the entities we studied, this dip is only observed in the Netherlands.

Green only open access

Solely based on these data, the role of green only open access seems modest. For the general universities, 'green only' accounts for 7% of all open access publications (4% of all publications, articles & reviews only), for the mixed and technical university, this is 3% and 1%, respectively. As noted before, this is not representative for the total level of green open access, as it only includes papers not also available as some form of gold OA. It also excludes author versions prior to peer review (e.g. all papers shared as preprint) The higher levels of green OA observed for the general universities in our samples might be explained by the coverage of PubMed Central by oaDOI. Whether the institutional repositories of all universities in our sample are harvested by oaDOI was unsure at the time of writing.

Bronze open access

The proportion of bronze open access increases for older articles, to the extent that for the general and mixed universities, more articles published prior to 2014-2015 are available through bronze open access than through gold or hybrid, respectively. Even for 2016, bronze accounts for 18-32% of all OA for the general and mixed universities. The role of bronze OA is more limited for the technical university, but even there, 11% of open access publications from 2016 are available through bronze OA, a share that increases for older articles. While bronze here includes all articles in non-DOAJ journals that do not have an open license, the increase in availability of bronze for older articles suggests that a considerable part of this is delayed open access - journals making (part of) their content open access after a certain period of time. As observed by Laakso and Björk (2013), the majority of articles from delayed OA journals become open between 6 months and a year after publication, with a long tail extending to 5 years after publication. The large presence of bronze in the open access landscape is important in the context of open access policies, an issue that has also been noted by Piwowar et al. (2017). Will governments and institutions include bronze OA in determining open access levels? Will funders and institutions accept bronze OA as fulfillment of OA mandates, despite the fact that these publications have no open license, and that they are seldomly OA immediately on acceptance or publication? Will the availability of delayed OA journals influence author's decisions to publish in gold/hybrid OA journals?

Our data also that the drop in open access levels in 2017 compared to 2016 and before (a consistent feature in all open access data retrieved by us from Web of Science) is primarily due to a decrease in bronze OA. This confirms the idea that the drop is largely a result of delayed OA. The 2017 drop is more pronounced for the general universities than for the technical university, with the mixed university occupying a middle position. The difference is proportionate to the overall share of bronze OA in the universities' output. So, can the levels of bronze OA (and delayed OA in particular) be explained by the disciplinary composition of universities' output?

Field metals

We already saw that there are marked differences between OA levels of the various fields (Fig. 3a-d). To get a better understanding of the source of observed differences between different types of universities, and test our assumption that these are mainly due to the research fields covered, we explored the relative contributions of main research disciplines (Life Sciences & Biomedicine, Physical Sciences & Technology, Social Sciences, Arts & Humanities) to pure gold, hybrid and bronze open access in 2016 research output (articles & reviews) for the four universities in our sample. Information on primary research discipline for each article was taken from the field Research area in Web of Science.

Looking at all articles & reviews classified as 'gold open access' in Web of Science (= 'all gold': pure gold + hybrid + bronze, Table 1A-B), we see that for the general and mixed universities, two-thirds of these are coming from Life Sciences & Biomedicine and a quarter from Physical Sciences & Technology. For the technical university, the situation is reversed, with three quarters of total publisher-mediated open access attributable to Physical Sciences & Technology. Somewhat surprisingly perhaps, for the universities in our sample, Social Sciences is responsible for only 4-8% of gold open access output in Web of Science (Table 1A-B).

Table 1 Distribution of 'all gold' OA articles+reviews (published in 2016) over main disciplines. A. absolute numbers, B. percentages.

A

	Leiden	Utrecht	Wageningen	Delft
Life Sciences & Biomedicine	1279	1792	884	169
Physical Sciences & Technology	553	743	348	604
Social Sciences	140	224	48	56
Arts & Humanities	25	21	5	10
(blank)	3	0	6	0
all disciplines	2000	2780	1291	839

B

	Leiden	Utrecht	Wageningen	Delft
Life Sciences & Biomedicine	64.0%	64.5%	68.5%	20.1%
Physical Sciences & Technology	27.7%	26.7%	27.0%	72.0%
Social Sciences	7.0%	8.1%	3.7%	6.7%
Arts & Humanities	1.3%	0.8%	0.4%	1.2%
all disciplines	99.9%	100.0%	99.5%	100.0%

It is important to realize that the percentages shown here are a reflection not only of the share of research output from these disciplines in general at these particular universities, but also of the levels of open access publishing within the respective disciplines. Especially for Social Sciences (with the exception of psychology) and for Arts & Humanities, open access levels are still very low, for various reasons, as illustrated earlier. Notwithstanding this caveat, these percentages give the needed context to consider to what extent observed differences *within* the broad category of 'all gold' open access articles (i.e. between levels of pure gold, hybrid and bronze open access) can be traced back to disciplinary differences (Table 2A-D).

Table 2 Relative shares of pure gold-hybrid-bronze OA articles+reviews (published in 2016) in main disciplines. A. Leiden, B. Utrecht, C. Wageningen, D. Delft.

A

Leiden	pure gold	hybrid	bronze	total
Life Sciences & Biomedicine	28.3%	37.3%	34.4%	100.0%
Physical Sciences & Technology	44.5%	16.3%	39.2%	100.0%
Social Sciences	23.0%	59.7%	17.3%	100.0%
all disciplines	32.6%	33.1%	34.3%	100.0%

B

Utrecht	pure gold	hybrid	bronze	total
Life Sciences & Biomedicine	33.2%	34.3%	32.5%	100.0%
Physical Sciences & Technology	60.8%	30.3%	8.9%	100.0%
Social Sciences	27.2%	60.3%	12.5%	100.0%
all disciplines	40.1%	35.4%	24.4%	100.0%

C

Wageningen	pure gold	hybrid	bronze	total
Life Sciences & Biomedicine	36.5%	39.9%	23.6%	100.0%
Physical Sciences & Technology	69.5%	24.4%	6.0%	100.0%
Social Sciences	10.6%	66.0%	23.4%	100.0%
all disciplines	44.4%	36.8%	18.8%	100.0%

D

Delft	pure gold	hybrid	bronze	total
Life Sciences & Biomedicine	27.2%	51.5%	21.3%	100.0%
Physical Sciences & Technology	36.9%	54.6%	8.5%	100.0%
Social Sciences	25.0%	67.9%	7.1%	100.0%
all disciplines	34.1%	54.7%	11.1%	100.0%

For 'all gold' open access articles from 2016 in Life Sciences & Biomedicine, about one third each is made available through pure gold, hybrid and bronze open access at the general universities in our sample. For the technical university, the share of hybrid open access is larger, which might be due to specific subdisciplines covered.

For Physical Sciences & Technology, interesting differences are observed between the universities studied, especially regarding bronze open access. For three of the universities, including the technical university, hybrid and gold open access are the predominant forms of publisher-mediated open access, with a very small role for bronze open access. However, for Leiden University, over one third of 'all gold' open access articles in Physical Sciences & Technology is bronze open access (n=217). Looking deeper into the subdisciplines within Physical Sciences & Technology, this can almost fully be attributed to the field of Astronomy & Astrophysics, for which 183 of the 187 articles in Leiden in 2016 are available through bronze open access. As discussed in the section on open access levels in Physical Sciences, a number of important journals in this field make their content available as bronze open access: the American Astronomical Society (bronze after 12 months) and the Royal Astronomical Society (bronze after 36 months). The prominent presence of Astronomy & Astrophysics as a field of study in Leiden (third largest in the world in terms of article output according to WoS data) thus at least partly explains the relatively high level of bronze open access at this university, also over time.

Despite relatively low absolute numbers (as reflected in the low share in publisher-mediated open access overall, see above), in Social Sciences, hybrid open access consistently emerges as the predominant mode of publisher-mediated open access in these data.

It should be mentioned that the relative share of especially bronze open access in the various disciplines might be different for older articles, depending on publisher policies on the time delay observed in delayed open access. However, the observed differences between main disciplines seem consistent with the findings of Piwowar et al. (2017), who looked at relative levels of pure gold, hybrid and bronze open access levels for articles published between 2009-2015: high shares of bronze for Life Sciences & Biomedicine fields, as well as for Earth & Space Sciences, which include Astronomy and Astrophysics.

These examples show the added value of considering (sub)disciplines represented at a specific institution and the breakdown of pure gold, hybrid and bronze open access in interpreting open access levels for that institution. This is especially important in assessing effectiveness of open access policies and identifying areas where efforts to increase open access levels could be most fruitful.

DISCUSSION

Generally speaking the potential explanations for different levels of OA could be:

- (1) Author behaviour: the degree to which authors choose to make their papers open either through green or gold open access routes. In turn that can depend on many things (acceptance of OA, availability of OA venues & infrastructure, availability of APC funding, APC levels etc.);
- (2) Publisher behaviour, with (a) the degree to which publishers decide to make full journal volumes or selected issues/papers open access and the timing of that and (b) the length of embargoes on green open access set by publishers.

Given high enough absolute figures, most *abrupt* changes in open access levels observed in our study are probably not due to researcher behaviour but to (changes in) publisher policies (including introduction and timing of delayed OA, green OA embargoes, specific OA deals such as SCOAP³) and database effects (changes in composition and inclusion policy of WoS). More gradual effects are more likely to be attributable to changing author behaviour, either driven by funder and institutional policies, changes in availability of open access publishing options (including financial options), and perhaps also general increasing awareness of the importance of OA among researchers making decision on where and how to publish.

Our analyses and the examples used demonstrate the added value of being able to combine information on (type of) open access with other variables such as institution, funder or country, especially when also being able to compare results by years of publication. The level of detail that is available with these data makes it possible to test hypotheses, check claims, corroborate reports from various stakeholders, evaluate open access policies and especially discover successful ones.

As the value of the OA filter in Web of Science has not been tested before, our methods and data need rigorous checks. What is its applicability for different fields and countries? Are its functionality and underlying data sufficient for all applications? The caveat of potential green OA underrepresentation remains especially important, both conceptually (what do

stakeholders consider acceptable types of open access and is there priority in those?), and statistically (how do our numbers compare with those using other sources (e.g. universities' own registrations)?

We realize that access to Web of Science is a privilege associated with affiliation with rich, mostly Western, universities. This also means that many of the analyses done here, particularly for non-Western countries and institutions, cannot be repeated or applied by the people directly involved or affected. We applaud oaDOI for their commitment to keeping their API free (ImpactStory 2017a) and are looking forward to collaborative initiatives to make information about research output from various entities openly available and reusable, enabling optimal (re)use of that information.

Regarding freely accessible alternatives to Web of Science: none of the three main candidates ([Google Scholar](#), [ScienceOpen](#) and [BASE](#), the Bielefeld Academic Search Engine) allow the easy analyses that are possible with the combination of WoS with oaDOI data (Table 3). ScienceOpen comes closest, but does not index green OA apart from some dedicated preprint servers. In addition, as mentioned on their site, their OA detection (based on information from creative commons license tagging in the article XML delivered by publishers and aggregators) is by no means comprehensive. There is one company that does carry out comprehensive OA-detection, but its products are not openly available: [1Science](#).

Table 3 Comparison of Web of Science with freely available databases for detailed determining of OA levels.

	WoS	BASE	Google Scholar	ScienceOpen
Filtering on OA	✓	✓	x	✓
Filtering by type of OA	✓	x	x	x
API	✓	✓	x	x
Exporting records with DOI info	✓ ¹	x	x	✓ ²
Filtering by document type	✓	✓ ³	x	✓
Filtering by country of affiliation	✓	✓ ⁴	x	✓ ⁵
Filtering by field	✓	x	x	✓
Filtering by affiliation	✓	x	x	✓ ⁵
Filtering by publisher	x	x	x	✓

¹ max. batch=500 ² max. batch=200 ³ articles include newspaper articles ⁴ filtering on country of repository is possible ⁵ free text search only

Suggestions for further research and improving the data availability and reliability

We have four suggestions for further research:

- (1) Testing hypotheses on the potential causes of the OA levels (e.g. those hinted at in this paper), including funder policies, institutional mandates, publisher policies, repository infrastructure, economic differences and publishing cultures.
- (2) Creating better insights into which publishers/journals are the main contributors to the various types of OA (per discipline group)
- (3) More detailed comparisons with other quantitative OA studies (esp. ScienceMetrix that informs EU policies)
- (4) Additional analyses using different or additional sources of bibliographical information, either globally or regionally or for specific language areas.
- (5) More detailed comparisons with self-reported (but mostly not openly available) levels of green and gold from universities

We have five recommendations to improve OA-tracking using WoS with the oaDOI data and increase reliability and transparency:

- (1) Ensure comprehensive harvesting of repositories by oaDOI and make the list of harvested repositories and exact ways of OA detection (including version detection) transparent and publicly available, perhaps including OpenDOAR IDs.
- (2) Include author-versions (preprints) in green OA data and provide a separate filter-category for that in WoS
- (3) Present filters for breakdown of gold (pure gold / hybrid / bronze) in WoS
- (4) Present all OA types of an article, with the ability to filter. That makes it possible to compare e.g. preprints with published versions and assess the development of various types of open access.

If these are addressed, the number of caveats issued at the outset could be reduced.

Finally, we have two wishes for implementations that could broaden potential for research and policy evaluations:

- (1) Offer publisher information as a searchable field in WoS, ideally as enhanced field
- (2) Make it possible in WoS to search on affiliation of corresponding authors
- (3) Make funding an enhanced field in WoS

CONCLUSIONS

Despite current limitations of the data sources, using WoS with the included oaDOI indications of open access produces useful results, especially for comparison of levels. The ability to combine OA information with data on fields, funding, languages, countries, institutions and topic facilitates many meaningful explorations. Adding more detailed information on OA segments by using the oaDOI API allows more detailed analysis and better suggestions for possible causes of the patterns observed.

The observed patterns of OA levels for the period studies reveal striking differences that could inform open access policies by the various stakeholders involved.

Further use of these data, including for hypothesis testing, would benefit from suggested improvements in the data provision. The low levels of additional value of green OA and the striking influence of bronze OA (confirming Piwowar et al. 2017) are very prominent general findings of this study, but the real value may be in what we can learn from the more specific patterns. For inspiration on successful policies it may be interesting to look at fields (e.g. astronomy/astrophysics, cell biology), countries (Tanzania, Netherlands, Latvia, Poland, Peru, Bangladesh) institutions (Harvard) and funders (Wellcome, NIH) that according to these data achieve high and/or increasing levels of open access.

REFERENCES

Archambault, É., Amyot, D., Deschamps, P., Nicol, A., Provencher, F., Rebut, L., & Roberge, G. (2014). Proportion of open access papers published in peer-reviewed journals at the European and world levels—1996–2013. European Commission. Retrieved from http://science-metrix.com/sites/default/files/science-metrix/publications/d_1.8_sm_ec_dg-rtd_proportion_oa_1996-2013_v11p.pdf.

Bielefeld University Library (2018) About BASE: Content sources. [Online, accessed at 20180107]. Available at: https://web.archive.org/web/20180107204555/https://www.base-search.net/about/en/about_sources.php.

Carvalho Neto, S., Willinsky, J. & Alperin, J. P. (2016). Measuring, rating, supporting, and strengthening open access scholarly publishing in Brazil. Education Policy Analysis Archives, 24(54). <http://dx.doi.org/10.14507/epaa.24.2391>.

Clarivate Analytics (2016a) An easier way to find open access. [Online, accessed at 20180101]. Available at: <https://web.archive.org/web/20180101184225/https://clarivate.com/blog/science-research-connect/an-easier-way-to-find-open-access/>.

Clarivate Analytics (2016b) Web of science core collection descriptive document. [Online, accessed at 20180101]. Available at: http://clarivate.libguides.com/ld.php?content_id=29366073.

Clarivate Analytics (2017a) About Web of Science Core Collection [Online, accessed at 20171220]. Available at: https://web.archive.org/web/20171220203111/http://images.webofknowledge.com/WOKRS527R13/help/WOS/hp_whatsnew_wos.html.

Clarivate Analytics (2017b) The Future of Open Access Discovery is NOW! [Online, accessed at 20171229]. Available at: <https://web.archive.org/web/20171229210910/http://info.clarivate.com/openaccess>.

Clarivate Analytics (2017c) Research Areas (Categories / Classification). [Online, accessed at 20171230. Available at: https://web.archive.org/web/20171230230106/http://images.webofknowledge.com/WOKRS527R13/help/WOS/hp_research_areas_easca.html.

Clarivate Analytics (2018) Master Journal List. [Online, accessed at 20180106. Available at: <http://mjl.clarivate.com/cgi-bin/jrnlst/jlresults.cgi?PC=EX>.

Directory of Open Access Repositories (2018) Open Access Repository Operational Statuses - Worldwide. [Online, accessed at 20180106]. Available at: <https://web.archive.org/web/20180106211029/http://www.opendoar.org/onechart.php?clID=&ctID=&rtID=&clID=&IID=&potID=&rSoftWareName=&search=&groupby=os.osHeading&orderby=Tally%20DESC&charttype=pie&width=600&height=300&caption=Open%20Access%20Repository%20Operational%20Statuses%20-%20Worldwide>.

Impactstory (2017a) oaDOI FAQ. [Online, accessed at 20180108]. Available at: <https://web.archive.org/web/20180108192808/https://oadoi.org/faq>.

Impactstory (2017b) oaDOI API v2 documentation. [Online, accessed at 20180108]. Available at: <https://web.archive.org/web/20180108232444/https://oadoi.org/api/v2>.

Johnson, R. et al. (2017) Towards a Competitive and Sustainable OA Market in Europe - A Study of the Open Access Market and Policy Environment. [Online]. Available at: [http://eprints.whiterose.ac.uk/114081/1/OA%20market%20report%20\(Final%2013%20March%202017\).pdf](http://eprints.whiterose.ac.uk/114081/1/OA%20market%20report%20(Final%2013%20March%202017).pdf).

Kieńć, W. (2017) Authors from the periphery countries choose open access more often. *Learned Publishing*, 30: 125–131. <https://dx.doi.org/10.1002/leap.1093>. Also available as preprint: <https://dx.doi.org/10.17605/OSF.IO/JU4RF>.

Kramer, B. (2018) Web_of_Science_OADOI_gold - R-script. v1.0 [Software]. Available at: <https://doi.org/10.5281/zenodo.1137861>

Kramer, B. & Bosman, J. (2018) Data from: Open access levels: a quantitative exploration using Web of Science and oaDOI data. [Dataset]. Available at: <https://doi.org/10.5281/zenodo.1143707>

Kunzmann, M. & Reckling, F. (2017) Austrian Science Fund (FWF) Open Access Compliance Monitoring 2016. [Dataset]. Available at: <https://doi.org/10.5281/zenodo.811924>.

Ministry of Education and Science of Latvia (2016), Latvian European Research Area Roadmap 2016-2020. [Online, access at 20180105]. Available at: http://www.izm.gov.lv/images/starptautiska_sad/Eiropas_P%C4%93tniec%C4%ABbas_telpaocw/Latvian_ERA_Roadmap_2016_-2020.pdf.

OCW (2013) Letter to parliament: Open Access to publications [Online, accessed at 20180106]. Available at: <https://www.rijksoverheid.nl/documenten/kamerstukken/2013/11/15/kamerbrief-over-open-access-van-publicaties>. English version: <https://www.government.nl/government/members-of-cabinet/sander-dekker/documents/parliamentary-documents/2014/01/21/open-access-to-publications>.

OCW (2017). National Plan Open Science. Available at: <http://doi.org/10.4233/uuid:9e9fa82e-06c1-4d0d-9e20-5620259a6c65>.

Openaccess.nl (2017) Publisher deals. [Online, accessed at 20180106]. Available at: <http://openaccess.nl/en/in-the-netherlands/publisher-deals>.

Piwowar H, Priem J, Larivière V, Alperin JP, Matthias L, Norlander B, Farley A, West J, Haustein S. (2017) The State of OA: A large-scale analysis of the prevalence and impact of Open Access articles. PeerJ Preprints 5:e3119v1 <https://doi.org/10.7287/peerj.preprints.3119v1>.

Priem, J. (2017) We found much IR version metadata was inaccurate. [Tweet. Online, Accessed 20180108]. Available at: <https://twitter.com/jasonpriem/status/940399805431472129>.

Priem, J. (2017) Or I should say *most* of it is updated weekly. [Tweet. Online, Accessed 20180108]. Available at: <https://twitter.com/jasonpriem/status/947181430106451973>.

Research Information (2017) Monitoring the transition to open access. December 2017. [Online, accessed at 20180103]. Available at: <http://www.universitiesuk.ac.uk/policy-and-analysis/reports/Documents/2017/monitoring-transition-open-access-2017.pdf>.

Rimmert C, Bruns A, Lenke C, Taubert NC. (2017) ISSN-Matching of OA Journals (ISSN-GOLD-OA) 2.0. Bielefeld University. <https://doi.org/10.4119/unibi/2913654>.

RIN (2015) Monitoring the transition to open access. A report for the Universities UK open access co-ordination group. [Online, accessed at 20180103] Available at: <https://www.acu.ac.uk/research-information-network/monitoring-transition-to-open-access>.

ScienceMatrix (2016) ERA progress report 2016. ScienceMetrix Study, Data gathering and information for the 2016 ERA monitoring – Technical Report. [Online, accessed at 20180105]. Available at: http://ec.europa.eu/research/era/pdf/era_progress_report2016/era_progress_report_2016_technical_report.pdf.

ULiège Library Network (2018) ORBi: general statistics. [Online, accessed at 20180108]. Available at: <http://orbi.ulg.ac.be/stats-chart?locale=en&type=global>.

United Nations (2015) Sustainable development goals. [Online, accessed at 20180102]. Available at: <http://www.un.org/sustainabledevelopment/sustainable-development-goals>.

Université de Liège (2017) 10 years of Open Access at ULiège [Online, accessed at 20180103]. Available at: https://web.archive.org/save/https://www.uliege.be/cms/c_9479183/en/10-years-of-open-access-at-uliege.

VSNU (2017) Open access developments in the Netherlands. Newsletter open access 34. [Online, accessed at 20180107]. Available at: <https://us7.campaign-archive.com/?e=&u=6e6efd9c9c87905323ae817ec&id=4553fa5fe0>.

Wikidata (2017) Wikiproject Zika Corpus. [Online, accessed at 20180107]. Available at: https://www.wikidata.org/wiki/Wikidata:WikiProject_Zika_Corpus.

ADDITIONAL INFORMATION

Competing Interests

Jeroen Bosman and Bianca Kramer are both affiliated to Utrecht University Library which supports and promotes open access and runs its own repository. They also run the project [101 Innovations in Scholarly Communication](#) and as such, are also involved in the execution of the National Plan Open Science of the Netherlands.

Author Contributions

Jeroen Bosman conceived and designed the experiments, collected the data, analyzed the data, wrote the paper, prepared figures and/or tables, reviewed drafts of the paper. Bianca Kramer conceived and designed the experiments, collected the data, analyzed the data, contributed analysis tools, wrote the paper, prepared figures and/or tables, reviewed drafts of the paper.

Funding

The authors received no funding for this work.

Acknowledgements

The authors wish to thank Jeroen Sondervan for extended conversations on the backgrounds of open access development and Jason Priem and Michael Habib for helpful Twitter exchanges on aspects of the data supplied by oaDOI and Clarivate Analytics.