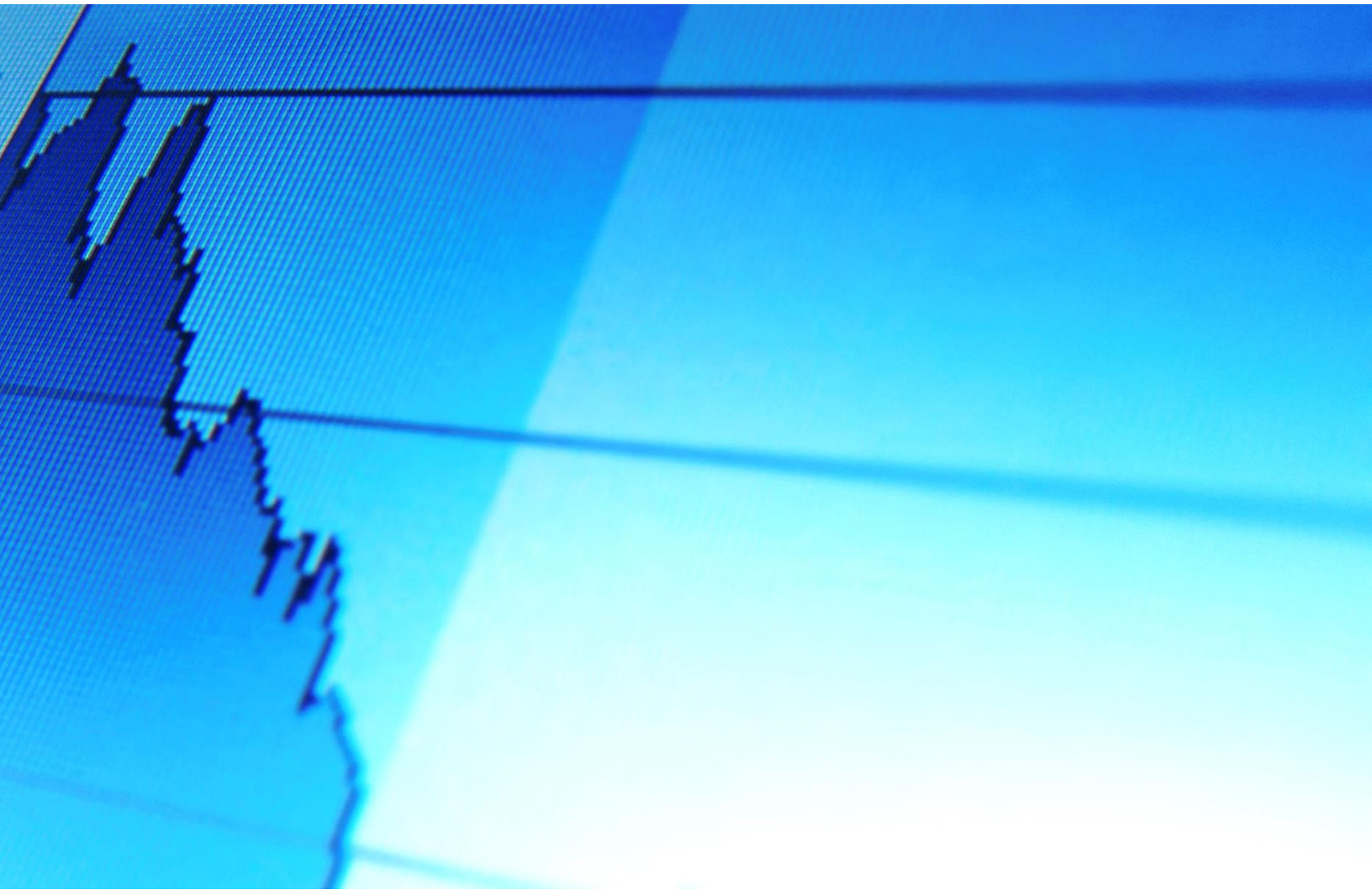
Science-Metrix

Analytical Support for Bibliometrics Indicators

Open access availability of scientific publications



Science-Metrix

Analytical Support for Bibliometrics Indicators

Open access availability of scientific publications*

Final Report

January 2018

By:

Science-Metrix Inc.
1335 Mont-Royal E. ▪ Montréal ▪ Québec ▪ Canada ▪ H2J 1Y6
1.514.495.6505 ▪ 1.800.994.4761
info@science-metrix.com ▪ www.science-metrix.com

*This work was funded by the National Science Foundation's (NSF) National Center for Science and Engineering Statistics (NCSES). Any opinions, findings, conclusions or recommendations expressed in this report do not necessarily reflect the views of NCSES or the NSF. The analysis for this research was conducted by SRI International on behalf of NSF's NCSES under contract number NSFDACS1063289.

Contents

Tables

Figures

Abstract

The adoption of national, regional and institutional policies to promote free access to new scholarly knowledge created with the help of public funds has driven the growth of open access (OA). In recent years, the level of availability has reached a tipping point, whereby at least half of the articles published become available in open access within 12 to 18 months of their publication.[1] Much needs to be learned about open access availability considering the rapid growth of this phenomenon. For example, how effective are mandates? Is open access growing because of the efforts of researchers themselves, such as in physics where preprints have been widely circulating for more than two decades? Or is the fast pace of change instead due to research funders or to institutional mandates? Answering these policy-relevant questions requires the use of robust measurement protocols. This report provides insight on very large-scale, quasi-population-level measurement and outlines current challenges and possibilities.

As with bibliometrics, where databases originally designed for bibliographic searching are being characterized and their data curated for bibliometric measurement, there is no dedicated source of data for open access measurement. This report compares established commercial databases—namely, the Web of Science and Scopus—with a bibliographic database that has been produced with the goal of facilitating the retrieval of *gold* and *green*[2] open access articles published in peer-reviewed journals. In addition to examining the strengths and limitations of large-scale measurement, this report performs a number of measures, particularly at the country and academic-field levels. It also examines the question of whether articles available in open access are more highly cited that those available strictly with a subscription.

The evidence presented in this report shows that at least two-thirds of the articles published between 2011 and 2014 and having at least one U.S. author can be downloaded for free as of August 2016. In the case of Brazil, the proportion reaches 75%. More broadly, the vast majority of the large scholarly publishing countries have more than 50% of their articles published from 2010 to 2014 freely available for download in gold and/or green gratis open access.

Examining the availability of articles by domains of scholarly activity shows that health sciences has the most articles available for free (at least 59% of the articles published in 2014 could be read for free in 2016), followed by the natural sciences (55%), applied sciences (47%), economic and social sciences (44%), and arts and humanities (24%). This is in part a reflection of the average number of authors on articles: the more authors on an article, the greater the probability that one of them will have funds to pay an article processing charge (for non-free gold OA) and that one author will take the time to archive the article on the public Internet.

Whereas current data suggests that gold OA is prevalent in health sciences, green dominates the natural sciences, applied sciences, and economic and social sciences. In the humanities, green and gold are more or less on the same level. Note that the level of undetermined OA type is high for all fields, and

[1] Archambault, É. et al. (2013). *Proportion of open access peer-reviewed papers at the European and world levels—2004–2011*. RTD-B6-PP-2011-2: Study to develop a set of indicators to measure open access. Montréal, Canada: Prepared for the European Commission Directorate-General for Research and Innovation. Retrieved from http://www.science-metrix.com/pdf/SM_EC_OA_Availability_2004-2011.pdf

[2] Gold open access involves the full text of an article being made available by its publishers, and green open access involves the full text being made available by parties other than the publishers. These terms are examined in detail in Section 2.1 of this report.

consequently these results by OA type can only be seen as an initial investigation of the question. This is linked to the challenge of attributing an OA type considering the hundreds of thousands of sources and the multitude of languages used on the Internet.

There is evidence that articles available in green OA are overall the most highly cited. This would be due to two phenomena being combined. Strictly green articles, meaning they are not otherwise made available by the publishers to the public, are published in journals that were established, generally speaking, a longer while ago compared to gold journals, which are a more recent phenomenon. As a result, these articles benefit from the high level of citedness of articles published in established, recognized journals. In addition, they benefit from a wider diffusion than non-OA articles, and therefore are more readily available. Articles published in gold journals are less likely to benefit from the reputation of the journals in which they are published, as the majority of gold journals were only established a few years ago. There are several factors influencing impact, and carefully crafted studies are necessary to determine the returns on investment achieved through different models of scholarly communication.

The measurement of the impact of open access on citations presented in this report is indicative rather than conclusive, being made as it is on millions of articles but lacking careful control of influencing factors. A similar conclusion can also be drawn on all the measures presented here. Bibliometrics is a complex science because it builds on data designed for purposes other than measuring. As a result, it is always necessary to characterize and curate data carefully, to study the results equally carefully, and to test factors that are influencing results to determine whether they are artifacts due to measurement protocols (including the source of data) or factors that are truly affecting the research system and the way results are being diffused. Open access is particularly challenging to measure. Whereas there are probably 5,000 to 10,000 publishers worldwide, there are millions of researchers potentially contributing their articles in green open access, and they do this in hundreds of thousands of places. The level of complexity is at least 10-fold greater, and so is the level of noise in the data. Moreover, measuring open access is a very recent activity, whereas earlier efforts in bibliometrics can be traced back to the 1920s. Hence, measurements such as those presented in this report are just the beginning and do not represent the final word on this complex and revolutionary transformation of the mode of access to research results.

1 Introduction

Given the adoption of national, regional and institutional policies to promote free access to scientific knowledge subsidized with public money, open access can be expected to contribute to the transformation of scholarly publishing.[3] Many countries, research funding agencies and research organizations have promulgated OA mandates and proposed policies to increase the availability of scholarly articles in open access.[4]

For example, knowledge circulation is a key policy aspect of the European Commission and the European Research Area, which see a need to give access to and preserve scientific information, and to promote open access to scientific publications and research data.[5] In pursuit of this, the European Commission has carried out significant work through pilot projects in two of its framework programs, FP7 and Horizon 2020 (H2020).[6] The FP7 pilot focused on open access to publications, which has become an underlying principle in H2020. An Open Research Data Pilot was launched as part of H2020 and was recently extended to cover all thematic areas of H2020, while ensuring opt-out possibilities for issues such as privacy, intellectual property rights (IPR) or national security concerns.

In the United States, the 2013 Office of Science and Technology Policy (OSTP) memorandum on "Increasing Access to the Results of Federally Funded Scientific Research" was put into effect[7] as the basis for the OA policy of the NSB, National Institutes of Health (NIH) and other federal agencies.

This memorandum states that

> The Office of Science and Technology Policy (OSTP) hereby directs each Federal agency with over $100 million in annual conduct of research and development expenditures to develop a plan to support increased public access to the results of research funded by the Federal Government. This includes any results published in peer-reviewed scholarly publications that are based on research that directly arises from Federal funds.

In this respect, it is relevant to put in place tools to measure the growth of open access. Other measures that would be relevant include estimating the contribution of mandates compared to other factors such as an innate motivation of researchers to share and diffuse their work. In both cases, it is also relevant to examine the means by which researchers make their articles openly available (e.g., institutional repositories, thematic repositories, academic social networks).

More specifically, the measurement protocol developed in the present report addresses the following questions:

- Is open access gaining ground? The present report examines the proportion of articles that are published in peer-reviewed/quality-controlled scholarly journals and listed in the reference databases

[3] Archambault et al., *Proportion of open access peer-reviewed papers at the European and world levels—2004–2011.*

[4] See Registry of Open Access Repository Mandates and Policies (ROARMAP): https://roarmap.eprints.org/

[5] Archambault et al., *Proportion of open access peer-reviewed papers at the European and world levels—2004–2011.*

[6] Archambault et al., *Proportion of open access peer-reviewed papers at the European and world levels—2004–2011.*

[7] https://obamawhitehouse.archives.gov/sites/default/files/microsites/ostp/ostp_public_access_memo_2013.pdf

(the WoS and Scopus) and that are also available in open access. The report examines how the evolution of OA availability can be studied.

- How does the current situation in terms of OA availability in the United States compare to that observed in other countries? The report critically examines the potential limitations in retrieving OA publications from the various Web repositories—namely, in the case of the 1science database, those articles published in other countries, particularly those using languages other than English and using character sets other than Latin in their text and metadata. Preliminary results are presented using a simple calibration.
- Given the policy goals of increasing access to publicly funded research, the report examines the level of citation received by millions of open access and non-open access articles to determine whether open access articles are more highly cited, and points to the challenges of studying the evolution of impact over time.

This study uses a measurement protocol based on a quasi-population-level approach—that is, one where all items would be counted. A definition of open access is discussed, including consideration of its differences from that promulgated in the Budapest Open Access Initiative. The central definition in the present study draws from the suggestion of Peter Suber to stress the *gratis* character of the access rather than considering free access in addition to the removal of rights on scholarly articles.

Three databases are used to measure the availability of articles in open access: the 1science database (by 1science, a sister company to Science-Metrix), Scopus (by Elsevier), and the Web of Science (the WoS, by Clarivate Analytics). The 1science database is a comprehensive bibliographic database, not unlike Scopus and the WoS, with the key difference being that it comprises metadata and hyperlinks to gratis OA articles published in peer-reviewed journals. Scopus and the WoS are the two most comprehensive bibliographic databases for scholarly and research publications, especially for serials, and to a certain extent for conference proceedings and monographs. For the purposes of this study, the 1science database needs to be used together with Scopus or the WoS because it only includes a record if an article is available in OA. Scopus and WoS are therefore used to perform quasi-population-level measurement and the 1science database is used to determine which of these millions of articles are available for free. Overall, producing statistics with the 1science database showed that Scopus and the WoS yield relatively similar findings regarding open access availability: 56% open access in the WoS and 52% in Scopus (availability measured in 2016 for articles published in 2013 using conservatively calibrated measures).

Given the present form of the databases used here and the matching algorithms used to link articles between these sources of data, the protocol used in this study presents more important limits for articles published in languages other than English and using character sets other than Latin. There are, for instance, more significant challenges measuring articles with Chinese authors compared to articles with U.S. authors. The gap is more pronounced when using the 1science database in combination with Scopus (56% of Chinese-authored OA articles found, against 74% for U.S. authored articles) than in combination with the WoS (73% against 80%). This is the result of Scopus covering more Chinese journals than the WoS and more of these journals not currently being whitelisted in the 1science database.

The protocol used in this study provides a generally similar underestimation of open access levels across scientific disciplines, retrieving about or close to 80% of the open access publications when either the

Scopus or WoS databases are used in conjunction with the 1science database. The only exception is the domain of arts and humanities, where the 1science database currently has 67% of the articles that are indexed in the WoS and that can also be discovered somewhere on the Internet (excluding dark open access sites containing mostly illegally obtained scholarly articles) and downloaded in an unencumbered manner.

Scholarly communication is changing rapidly, and this report shows that because of the ever-changing stock of material available on the Web, measuring temporal changes and growth in open access presents specific challenges. This is due to the backfilling of older publications in repositories, embargoes on older publications coming to an end, and publishers changing their access policy retrospectively and making older content openly available. Repeated snapshots and longitudinal studies of the state of open access are needed to address these issues properly.

Section 2 presents the measurement protocol used in the present study. This starts with a definition of open access and continues with the specific methods used in the study, followed by the characterization and a calibration of the measurement instrument. OA measurements are conducted in Section 3, and Section 4 concludes the report.

2 Methods

2.1 Definition of open access

The openness of scholarly articles varies greatly and reflects many factors, such as who the owners of the rights are, what rights for articles are provided by different types of licenses, where the articles are stored and how discoverable they are, to name just a few variables. It is not surprising in this context to find that there is no consensus on the definition of open access.

Three important meetings of the open access community, held successively in Budapest, Bethesda and Berlin, have given rise to a relatively strict definition of open access that is sometimes referred to as the BBB definition. The introductory paragraph of the original declaration of the Budapest Open Access Initiative (BOAI) is interesting as it provides context, justification, prescription and intended effects:

> An old tradition and a new technology have converged to make possible an unprecedented public good. The old tradition is the willingness of scientists and scholars to publish the fruits of their research in scholarly journals without payment, for the sake of inquiry and knowledge. The new technology is the internet. The public good they make possible is the world-wide electronic distribution of the peer-reviewed journal literature and completely free and unrestricted access to it by all scientists, scholars, teachers, students, and other curious minds. Removing access barriers to this literature will accelerate research, enrich education, share the learning of the rich with the poor and the poor with the rich, make this literature as useful as it can be, and lay the foundation for uniting humanity in a common intellectual conversation and quest for knowledge.[8]

This definition can be said to be strict because of the requirement that articles not only be free, but also be provided via unrestricted access. In its core definition, the BOAI enumerated a series of requirements for what can be considered as unrestricted open access:

> By "open access" to this literature, we mean its free availability on the public internet, permitting any users to read, download, copy, distribute, print, search, or link to the full texts of these articles, crawl them for indexing, pass them as data to software, or use them for any other lawful purpose, without financial, legal, or technical barriers other than those inseparable from gaining access to the internet itself. The only constraint on reproduction and distribution, and the only role for copyright in this domain, should be to give authors control over the integrity of their work and the right to be properly acknowledged and cited.

There has been constant debate, still ongoing today,[9] on whether this definition may have been too strict and that a more permissive approach could have proven more useful in providing a stepping stone toward achieving this more extensive set of requirements. This led Peter Suber, one of the original signatories of the BOAI, to suggest the following distinction between two types of open access:

> The term "open access" is now widely used in at least two senses. For some, "OA" literature is digital, online, and free of charge. It removes price barriers but not permission barriers. For others, "OA" literature is digital, online, free of charge, and free of unnecessary copyright and licensing restrictions. It removes both price barriers and permission barriers. It allows reuse rights which exceed fair use. There are two good reasons why our central term became ambiguous. Most of our success stories deliver OA in the first

[8] http://www.budapestopenaccessinitiative.org/read

[9] http://poynder.blogspot.co.uk/2017/01/the-nih-public-access-policy-triumph-of.html

> sense, while the major public statements from Budapest, Bethesda, and Berlin (together, the BBB definition of OA) describe OA in the second sense.
>
> I've decided to use the term "gratis OA" for the removal of price barriers alone and "libre OA" for the removal of price and at least some permission barriers. The new terms allow us to speak unambiguously about these two species of free online access.[10]

This report uses *gratis OA* as the operational definition of open access, and all the calibrated measures are based on this (the need for calibration is explained in the present section). It is also important to distinguish between different types of open access. In this respect, an important distinction is frequently drawn between two types of open access based on where the copies of the articles are located and who made them available. For instance, Björk and colleagues state that "green OA is defined as all freely accessible copies of articles, including different versions of said articles, which exist on other web locations than the original publisher's website."[11] This definition places gold open access strictly on publishers' websites. In practice, this does not always work, as some of the material is placed on other sites by publishers. Therefore, the 1science database uses a slightly different definition:

> Gold OA refers to papers made available for free by the publishers themselves, be it on their website (e.g., in fully gold OA journals on Springer Open and BioMedCentral, or as hybrid OA, that is, OA papers from otherwise paywalled journals on, for example, Springer's website) or on the site of an aggregator (e.g., SciELO, and also PubMedCentral, on which the majority of papers are archived by the publishers themselves).
>
> Green OA refers to papers made available for free by parties other than publishers, usually the authors themselves, who archive papers in institutional repositories, subject repositories such as arXiv, or commercial repositories such as ResearchGate.[12]

Note that articles might be hosted in more than one location and can be available through both the gold and green routes, meaning that the two categories are not mutually exclusive.

In the present study, articles are considered as being gratis OA if they are available on the public Internet (i.e., sites that don't require a registration) in full-text form and can be read and downloaded for free. One can also add that because the BOAI specifically mentioned "use them for any other lawful purpose," anonymous websites whose raison d'être and modus operandi are primarily to diffuse illegally obtained scientific literature are excluded from the present measurement. That is not to say that the world of open access can be described in a dichotomous manner: there are grey zones such as the personal Web pages of researchers, academic social networks and even some institutional repositories that contain a combination of articles fit for archiving in that specific place and others that should not appear there, or perhaps not in that form.

It is important to reflect on the stages of production and publication of the scholarly literature in order to understand open access to these scholarly articles. The first stage is the writing of articles, commonly known as *papers* in academia, which are then submitted to a scholarly journal for review. That first version of the

---

[10] http://sparcopen.org/our-work/gratis-and-libre-open-access/

[11] Björk, B.-C., Laakso, M., Welling, P., & Paetau, P. (2014). Anatomy of green open access. *Journal of the Association for Information Science and Technology*, *65*(2), 237–250. doi:10.1002/asi.22963

[12] Archambault, E., Côté, G., Struck, B., & Voorons, M. (2016). Research impact of paywalled versus open access papers. Science-Metrix & 1science. Retrieved from http://www.1science.com/oanumbr.html

paper is called the *preprint,* and open access can trace some of its roots back to the circulation of papers by authors to colleagues, sometimes before submission to journals (draft papers, which are not preprints per se). In physics, it became common to start circulating these preprints widely to colleagues, and that saw the advent of the arXiv preprint server in the early 1990s, which essentially became an archive of preprints (and currently also archives post-prints). The next stage of the publication process involves peer review or quality control assured by a journal editor—such as is the case in some fields, particularly in the social sciences and humanities where the practice coexists with peer review. Once a paper has been accepted for publication, and often changes have also been made by authors, it becomes known as the *post-print*. Finally, journal publishers do the page layout, in some cases provide value-added services such as proofing and—more rarely—improving tables and graphs, and they create what is known as the *version of record*, also sometimes known as the *publisher's PDF*.

The open access sphere has been made fairly complex due to the presence of these three embodiments of the papers. Although in many cases there are strict copyrights in place, and journal publishers own these rights to the articles,[13] most publishers have accepted there is a need for articles to be shared among researchers and they should be accessible to the public, who have supported research through their taxes. However, the rules are complex and vary greatly between publishers. Most publishers allow researchers to make their preprints available, sometimes after an embargo period. Many publishers also allow the final peer-reviewed version to be archived, and embargoes are somewhat more frequent in this case. Finally, there are also publishers that allow researchers to archive the version of record. What complicates this is that rules concern not only the state of an article, but also where it can be archived, and when.

Many publishers' rules can be found on the SHERPA/RoMEO website,[14] but there are also agreements that are not made public, which are the result of negotiations between an organization and a publisher. This means that in some cases a governmental organization has obtained the right to post the version of record of papers when an author of the paper is an employee of that organization, but a university with an author on that same paper cannot copy the paper to its institutional repository. Hence, it is currently impossible to determine which of the tens of millions of papers that are freely available on the Web are in a form acceptable to or tolerated by publishers. This is why the present report excludes those anonymous websites built on the basis of illegally obtained articles but includes sites where researchers have themselves decided to share their own papers, even if careful research would reveal that a portion of those papers may not respect embargoes or are not in the specific form of the paper their publisher permits to be shared. It is acknowledged that sharing is a tradition of the scientific community that continues in open access. Because of this, academic social networks are included here provided that their articles are available without registering first. Rather than seeing measurement efforts stall while attempting to perform the impossible task of judging whether these millions of papers are shared in a form acceptable to publishers, the stance here is that it is for researchers to share responsibly and for publishers to defend their rights if they consider papers shouldn't be shared. Hence, the approach proposed here is pragmatic and doesn't advocate for

[13] Authors often have to cede their rights to their papers if they want to publish in the most prestigious journals. The academic journal publishing industry presents a rather rare model, one where authors do not have rights and do not receive royalties. Compare this to newspapers where authors usually get paid for article writing, or to book and academic manual publishing where authors receive royalties.

[14] http://www.sherpa.ac.uk/romeo/search.php

either the author's or the publisher's view. It measures what is out there, available for the public to read and to download from the public Internet, and leaves the issue of the management of rights to be negotiated between stakeholders in the system.

As one can see, the bibliometrics of open access is substantially more complex than measuring only published papers as undertaken in traditional bibliometric statistics. That said, the operational definition based on gratis open access used in the present report presents clear boundaries, and it enables measurement to move forward while the rights issues continue to be negotiated as an important but separate issue from basic access to reading and downloading.

2.2 Approach to measurement of open access

Usually, studies aimed at measuring scholarly output require that many if not all relevant articles be counted. In practice, this means that articles are sampled from an authoritative source that claims to more or less exhaustively contain all relevant articles. This technique has traditionally been used to measure the availability of the scholarly literature in open access form. The ideal measurement would be at the entire population level; that is, all relevant articles would be measured to determine whether they are openly available are not. For the present study conducted for the NSF, it was decided to go one step further than previous studies on open access and to perform a *population-size* measurement.

Importantly, performing population-size measurement creates problems of its own. One of the problems is the absence of a bibliographic database that contains all the existing scholarly literature. There are comprehensive databases, however, that can be expected, by and large, to reflect the overall literature. These include the Web of Science, which is produced by Clarivate Analytics, and Scopus, produced by Elsevier. One could use these databases as a reference and verify how many articles can be found in OA.

While doing so, it is important to state that what is being measured is not the percentage of papers in open access overall, it is the percentage of articles in these databases that are available in OA form. The distinction is important for several reasons, including but not limited to the following:

1. Both the WoS[15] and Scopus[16] concentrate on the larger and the more highly cited journals.
2. They have a bias toward English-language journals.
3. They have a bias toward Latin-alphabet journals.
4. They have a bias toward journals published in Western countries.
5. The WoS has a bias toward the natural and health sciences, with less coverage of the arts and humanities.
6. They have a potential bias toward subscription-based journals, with proportionately less coverage of freely available, gold access journals.

[15] http://wokinfo.com/essays/journal-selection-process/

[16] https://www.elsevier.com/solutions/scopus/content/content-policy-and-selection

The bias toward subscription-based journals is worth articulating in more detail here. Ulrich's Periodicals Dictionary is a valuable benchmark and lists approximately 35,000 peer-reviewed scholarly journals worldwide (as of the end of 2014).[17] DOAJ lists approximately 10,000 gold OA journals,[18] meaning that just under 30% of all journals are available in gold OA. Scopus covers a total of about 27,000 journals, but less than 3,000 of them (about 10%) are gold OA journals. As for the WoS, of the 18,000 journals covered, only about 1,200 (less than 7%) are gold OA journals. These findings are not conclusive, and a full-scale study devoted to such a bias would be required to give a definitive answer to the question, but the coverage of Scopus and the WoS appears to have at least some bias in favor of subscription-based journals, which is an important consideration in the context of the present study.

In the present study, the whole population of articles in the WoS and Scopus databases are used as the denominator to calculate OA availability, so each article in these databases is considered as either available in gratis OA or not. The numerator is provided by matching these articles to a database developed by 1science,[19] which catalogues open access (OA) scholarly articles, across fields of research, inventorying both hyperlinks to full-text content and metadata relating to the publications themselves. The salient features of this database, for the context of the present study, are the following:

- Articles considered OA in the 1science database are those that are "digital, online, and free of charge," following the gratis OA definition outlined above. Furthermore, articles in the database must be available for download in an *unencumbered* fashion, meaning that they cannot be hidden behind logins, passwords, CAPTCHAs or other barriers—even though a human user can overcome such barriers manually and without monetary cost.[20] The definition of open access used to build the 1science database thus diverges slightly from the definition used for the present study.
- Articles can be posted in OA by either the publisher (gold OA) or by another party, such as the researchers involved in producing the publication, the institutions at which they work, and so forth (green OA).
- Anonymous sites providing mostly illegal access to articles are not crawled by the 1science harvester to document their contents. Although a certain portion of their OA content is made available legally, an exhaustive detailing of the full variety of rights ownership and licensing agreements is not feasible.
- The 1science database is still in a period of rapid expansion; over the course of 2016, the number of articles indexed grew from 7.6 million to 23.2 million. It indexes scholarly material found on over 180,000 websites, though many sources have certainly not yet been discovered by the harvesting modules that populate the database.
- Like WoS and Scopus, 1science has a bias toward publications written using the Latin alphabet.

[17] Cited in Ware, M., & Mabe, M. (2015). *The STM report: An overview of scientific and scholarly journal publishing*. The Hague, The Netherlands: International Association of Scientific, Technical and Medical Publishers. Retrieved from http://www.stm-assoc.org/2015_02_20_STM_Report_2015.pdf

[18] Wohlgemuth, M., Rimmert, C., & Winterhager, M. (2016). ISSN-Matching of Gold OA Journals (ISSN-GOLD-OA). Bielefeld University. Retrieved from https://doi.org/10.4119/unibi/2906347

[19] http://www.1science.com

[20] For a discussion of encumberment specifically, see Section 2.3.

- The primary purpose of the 1science database is to provide access to scholarly material available in OA, rather than to be used as a basis for bibliometric measurements; it is also not intentionally designed to optimize overlap with the content of WoS or Scopus databases.

As noted above, there is a divergence between the definition of OA embodied in the 1science database and the one at issue in the present study. Furthermore, the databases being used all have certain known biases, and there are potential problems that can occur when matching data between the WoS, Scopus and 1science databases. Accordingly, measurements of OA are calibrated, to provide a more accurate reflection of the underlying reality.

2.3 Method used for instrument calibration

It is possible to correct for all these limits by calibrating the measurement system against an alternate system, or *reference system*. Limitations can be characterized with the help of two indicators, drawn from the realm of information science: *precision* and *recall*. The precision rate shows how much of the information captured by the measurement system is relevant in the reference system. The recall rate shows how much of the relevant information in the reference system is captured by the measurement system. If all the material captured by the measurement system is relevant, it has a precision of 100%; if all the relevant material is captured by the measurement system, it has a recall of 100%.

Four measures are necessary to assess retrieval precision and recall: true positive results, true negative results, false positive results and false negative results. True positive (*tp*) results are records that are identified as being available in open access by the instrument and that are indeed available according to verification. True negative (*tn*) results are records that are identified as not being available in open access by the instrument and that are indeed not available according to verification. Following the same logic, false positive (*fp*) results are records identified as being available by the instrument but that are not so according to verification, whereas false negative (*fn*) results are records identified as not being available by the instrument but that are in fact available according to verification.

These four metrics make it possible to compute retrieval precision and recall. Retrieval precision (i.e., % of publications identified as open access that are open access) is computed as follows:

$$\text{Retrieval precision} = \frac{tp}{tp + fp}$$

Recall (i.e., % of documents available online that could be retrieved by the instrument) is computed as follows:

$$\text{Recall} = \frac{tp}{tp + fn}$$

Tests performed at various times in the last year have shown that the 1science database's precision is quite high (typically 97% or greater). A high level of certainty is therefore associated with documents counted as OA by the instrument; that is, publications flagged as being available in open access are almost always available and are indeed published in peer-reviewed journals. In the few rare cases where they are not available, it is almost always because of broken Web links due to changes made to websites (e.g., a change from http to https in the URL).

Another type of error affecting precision would be that articles are tagged as belonging to a peer-reviewed journal when in fact they are not. Tests reveal this to be a rare occurrence. For a variety of reasons, it can happen that the wrong document is downloaded in place of the one that should be downloaded, which affects precision. For example, some institutional repositories point to PDF documents as if they were the full-text article, but instead they only contain the abstract of the article. Over time, the occurrence of this type of relatively uncommon error will be reduced further by the development of additional data validation techniques. Generally speaking, an instrument containing only the documents in the gold standard would have a precision of 100%.

Recall is affected by variables such as the quality of the metadata that accompany hyperlinks to full-text articles, the language of the website and metadata, and certainly the fact that not all sources of eligible materials are known to 1science yet. Non-Latin characters present challenges to the system and are more likely to yield matching errors. It is noteworthy that at the current stage of technological development, retrieving a publication in English is far easier than retrieving a publication in Mandarin (among other languages), which probably leads to a bias against China for the measurement of open access levels.

In the case of publications tagged by the 1science database as being available in open access, download links to the PDFs were used to test if the publications were indeed available online and to ensure that there were no mistakes in the matching process with the 1science database. Each PDF link was tested, and in the case of broken links—which can happen as the Web is constantly evolving—or of erroneous assignations made by the harvester, analysts tried to retrieve publications online following the manual searching process described below. If these publications could not be retrieved, they were flagged as not being OA.

With the measures outlined above, a calibration factor can be computed, and applying it to measurements taken by comparing 1science to WoS and Scopus can provide results better suited to the context of the present study. This calibration factor is computed as follows:

$$\text{Calibration factor} = \frac{tp + fn}{tp + fp}$$

In order to compare the measurement system to the reference system, a characterization of the reference system is required. To obtain this characterization and facilitate the calibration of measurements, the open access status of scientific publications was validated; random samples of 500 publications were checked manually online for each of the five domains of scholarly activities used in the Science-Metrix classification.[21] A set of 500 Chinese publications and another set of 500 U.S. publications were used to characterize the 1science database at the country level. Samples of 500 publications were used as they yielded a margin of error of about 4% (95% confidence interval), which was deemed precise enough in the context of this study. Larger samples would have yielded a more precise characterization, but considering that doubling sample sizes to reach 1,000 documents only reduced the margin of error by 1 percentage point, it was decided that the additional workload outweighed the benefit.

[21] Version 1.06. See http://www.science-metrix.com/en/classification

To perform these analyses, publications were randomly selected for each category stated above. Titles, first authors and publication years were used to search the Web for publications that were flagged as not available in open access in the 1science database. Analysts manually searched for these publications using the Google search engine, focusing on the 20 to 30 first results presented, after which search results were found to be generally unrelated to the documents being searched. Following searches, analysts could record their decision regarding the open access status of these publications.

Please note that this method is not perfect in itself. Some users are more proficient in advanced search functions, whereas other users are less proficient. Some will sift through a greater number of irrelevant search results looking for an item further down the list of search results, turning up a result that other users would not find, having abandoned their search earlier in the process. Furthermore, although our analysts are fluent in both English and French, retrieving publications online from sources covering a multitude of languages is challenging. For instance, while verifying samples, a few searches led to a specific Web page in Arabic that listed publications but never seemed to permit the retrieval of full documents. Although analysts took care in investigating these cases, considering current shortcomings in the machine translation of Web pages, they might have missed publications from some sources.

Many parameters must be stipulated to define a reference system in a methodologically consistent way—including the skills of users, cut-off points about the number of results to examine, and so forth—where these methodological decisions contribute significantly to what is being defined as open access or not. Google was selected as the main search engine for this exercise because of its efficiency in retrieving documents online, powered by underlying tools such as Google Scholar. It was also selected based on its status as the most widely used search engine worldwide, being used for 68% of all searches.[22] However, even Google does not have a perfect precision and recall relative to other given approaches. It remains a valuable benchmark, due to its ubiquitous usage and its power, but should not for these reasons be conflated with an "objective" measure of open access. Accessibility is a notion that is always defined relative to many parameters, including the user, the tool and the target.

2.4 Calibration of measurements at the country level: China and the United States

A series of characterization exercises were conducted to verify how well the 1science database measured OA availability at the country level and determine the calibration that could be used to obtain a measure closer to the definition of gratis OA used in the present study. Although it would be useful to perform this kind of characterization for every country, this was well beyond the scope of this project given the amount of work involved in manually searching papers in 500-article samples. Consequently, it was decided to focus on two countries of interest to the NSF: the United States and China. These countries are the world leaders in scientific output, with a combined participation in more than 40% of all scientific publications at the world level in 2014, as measured using both the WoS and Scopus databases. Furthermore, because of linguistic differences, selecting these two countries enabled the characterization of the 1science database for an English-speaking and a non-English-speaking country. Because China does not use the Latin

[22] http://www.digitaltrends.com/web/google-baidu-are-the-worlds-most-popular-search-engines/

alphabet, this choice could also be expected to yield and highlight specific problems in both harvesting and data coupling.

Table I presents the results of the manual characterization of 500-article samples performed for both countries using the WoS as the baseline. The first notable finding is that retrieval precision is almost the same in both cases, standing at 97% for China and 98% for the United States. This means that the 1science database is quite precise when it comes to identifying open access publications, only rarely being mistaken in identifying a publication as open access when it should not. Furthermore, the few cases where the 1science database erroneously identified a publication as open access were usually the result of broken or disappearing links, which means that these publications were likely available when the Web was originally scanned, but have since disappeared or been moved. These broken links highlight the transient nature of open access, with the Web constantly evolving and changing. In some cases, the errors could be due to a system malfunction, as it is extremely challenging to always provide a direct link to papers given that about 180,000 sites provide content used in the 1science database.

Although retrieval precision is similar for both countries, there is a notable difference regarding recall of open access publications. The 1science database contains 80% of the U.S. papers defined as being open access in the present study, whereas this was the case for only 73% of Chinese open access publications. Several phenomena are simultaneously at play. Firstly, there are several more Chinese peer-reviewed journals that have not yet been tagged as peer reviewed by 1science, which would mean that even if material is collected, it doesn't make its way to the measurement database. This is because the search for an authoritative list of Chinese peer-reviewed journals has proven elusive, and this is an essential step in the 1science quality control procedure. In the absence of tangible information that a journal is refereed or peer reviewed, none of its data are accounted for. An additional factor is that it is more likely there are problems with the matching of metadata between Scopus or the WoS and the 1science database in Mandarin than in English. Finally, it is more challenging generally to discover full-text articles on Chinese sites compared to those using English.

Table I Calibration of open access status of publications, by country affiliation in the WoS (2006–2010)

Category	True positives (tp)	True negatives (tn)	False positives (fp)	False negatives (fn)	Retrieval precision	Recall	Adjustment
China (WoS)	174	254	6	66	97%	73%	1.33
United States (WoS)	283	140	7	70	98%	80%	1.22

Source: Prepared by Science-Metrix using the Web of Science (Clarivate Analytics) and the 1science database[23]

The combined effect of the factors stated above leads to a slightly stronger divergence from measured open access levels in the 1science database for China than for the United States. Based on the parameters presented above, the calibration for the United States should be 1.22, compared to 1.33 for China. Based

[23] For tables I, II and III, publications were limited to the 2006–2010 period to avoid including publications that were still covered by an embargo period for open access. Color gradient ranges from dark red for the lowest possible value (i.e., 0%) to white for values on par with the average and dark green for the highest possible value (i.e., 100%). Random samples of 500 publications were used for each academic domain or country.

on these findings, gratis open access measurements for the United States in 2013 would stand at 67% (55% prior to calibration) and 51% for China (38%).

Using the Scopus database (Table II), retrieval precision is observed to be similar for both countries, and the results are almost identical to those obtained using the WoS (96% for China, 98% for the United States). However, regarding recall, results point toward a larger divergence for China using Scopus, as only 56% of Chinese publications that were manually detected as being available in open access were identified as OA in the 1science database. This result is substantially lower than that in the WoS (73%). This is not the case for U.S. publications as the 1science database recall for Scopus is 74%, somewhat similar to but lower than the 80% mark observed for the 1science database recall for the WoS. Because of these lower recall levels for China, there is an important difference in its calibration factors (1.33 in the WoS, 1.71 in Scopus), whereas calibration factors for the United States are relatively similar in both databases (1.22 in the WoS, 1.33 in Scopus).

Table II Calibration of open access status of publications, by country affiliation in Scopus (2006–2010)

Category	True positives (tp)	True negatives (tn)	False positives (fp)	False negatives (fn)	Retrieval precision	Recall	Adjustment
China (Scopus)	100	318	4	78	96%	56%	1.71
United States (Scopus)	251	155	5	89	98%	74%	1.33

Source: Prepared by Science-Metrix using Scopus (Elsevier) and the 1science database

For the reasons explained above, the fact that Scopus comprises a greater variety of journals across regions and countries can possibly lower recall (as more of these journals may not yet have been whitelisted by 1science at the time of production). This could affect China more severely, as Chinese journals are more prevalent in Scopus than in the WoS.

Based on the adjustment factors presented above, open access measurements in 2013 should stand at about 68% for the United States (vs. 51% as measured in the 1science database) and 43% for China (vs. 25%). Although values vary quite substantially across databases for both countries, and more so for China, calibrated open access levels are relatively similar regardless of the database selected, especially for the United States (67% in the WoS for the United States against 68% in Scopus; 51% in WoS for China against 43% in Scopus).

2.5 Calibration of measurements at the domain level

Analyses similar to those presented at the country level in Section 2.4 were also prepared by academic domain to reveal the limits the 1science database might have at that level and to determine the best calibration to mitigate these limits. Table III presents precision and recall measures in addition to the adjustment required for calibration for the five main domains of the Science-Metrix classification (excluding the general category) using the WoS database.

Table III Calibration of open access status of publications, by academic domain in the WoS (2006–2010)

Category	True positives (tp)	True negatives (tn)	False positives (fp)	False negatives (fn)	Retrieval precision	Recall	Adjustment
Applied Sciences	198	255	3	44	99%	82%	1.20
Arts & Humanities	101	346	3	50	97%	67%	1.45
Economic & Social Sciences	213	213	6	68	97%	76%	1.28
Health Sciences	248	183	6	63	98%	80%	1.22
Natural Sciences	222	216	7	55	97%	80%	1.21

Source: Prepared by Science-Metrix using the Web of Science (Clarivate Analytics) and the 1science database

As was the case at the country level, retrieval precision is high and similar across domains, ranging from 97% to 99%. Recall is broadly uniform across disciplines, except for the arts and humanities. Recall values stand between 76% and 82% for natural sciences, applied sciences, economic and social sciences, and health sciences. As a consequence, calibration factors are relatively similar, ranging from 1.20 in applied sciences to 1.28 for economic and social sciences.

The domain for which recall stands out is arts and humanities. At 67%, it is about 12 percentage points below the recall value observed in other domains. Publications in this domain are more difficult to find online, as its journals are frequently smaller, with a more regional focus than journals in the natural sciences, for instance. Additionally, the metadata for humanities journals are frequently not made available in a manner conducive to effective harvesting. Furthermore, there is a higher proportion of non-English-language journals in the arts and humanities compared to other disciplines, and this probably contributes to the lower recall rate as well. Because of this lower recall rate, the calibration factor for arts and humanities is 1.45. Considering that the arts and humanities domain is not included in the Science and Engineering Indicators (SEI), this discrepancy would not represent a limitation if the SEI were to include open access measurements in the future.

One could expect additional variations in recall across scientific subfields. As an example, recall levels might not be the same for chemistry and biology, even though both fields are in the natural sciences domain.

2.6 The need to use calibrated results

As shown, the 1science database has a high level of precision but a relatively low recall. In this respect, given the level of development at the time this study was performed, it is safe to assume that open access levels as measured by the 1science database always represent an underestimation of the real measure. It is expected that this underestimation of open access levels affects publications involving authors from non-English-speaking countries more heavily given the composition of our team of analysts.

Considering the results of the characterization of the 1science database performed in this section, the use of a general calibration of 1.2 for all measures presented here appears to be required to obtain a truer estimate of gratis OA availability. This is a conservative calibration of measures taken by the 1science database. It coincides, by and large, with the lowest level of calibration observed, is in line with the calibration observed in applied sciences, and is slightly lower than that which should be used to more precisely assess OA in the United States when basing measures on the 1science database. This calibration of the instruments helps obtain a truer measure, while minimizing the risk of overestimating the proportion of OA in most measurements.

3 Calibrated open access measures using the Web of Science

Whereas the previous section of this report was dedicated to the presentation of the methodological approach and characterization of open access measurements using the 1science database, Scopus and the Web of Science, the present section of the report presents a succinct portrait of open access at the world level. Descriptive analyses presenting current trends in open access were prepared using the Web of Science database. This exercise was done using only one database to minimize duplication of effort, but the reader should note that the same exercise could also have been performed using Scopus or both databases. The section estimates open access availability at the country and world levels, as well as within specific scientific disciplines. Furthermore, data sources of open access, analyses regarding open access status (i.e., green and gold open access), and the citation advantage of open access are presented.

3.1 Stationary analysis of open access availability per year

Building time series to examine open access availability is challenging. One reason regular time series analyses cannot be performed readily is the presence of delays and embargoes affecting open access. Delays can result from the latency associated with self-archiving by researchers, but the main cause is the effect of publishers' embargo periods that ban the online archiving of articles until a certain period of time has elapsed. Another cause is journals that use paywalls (access by paid subscriptions only) for a set period before the articles become available for free. Because of these delays, open access availability curves show an inflection for the most recent years, as seen in Figure 1. Note that this measure examines papers available in the WoS and for which a free version could be found in the 1science database, and includes a calibration factor of 1.2 (all figures are augmented by 120%).

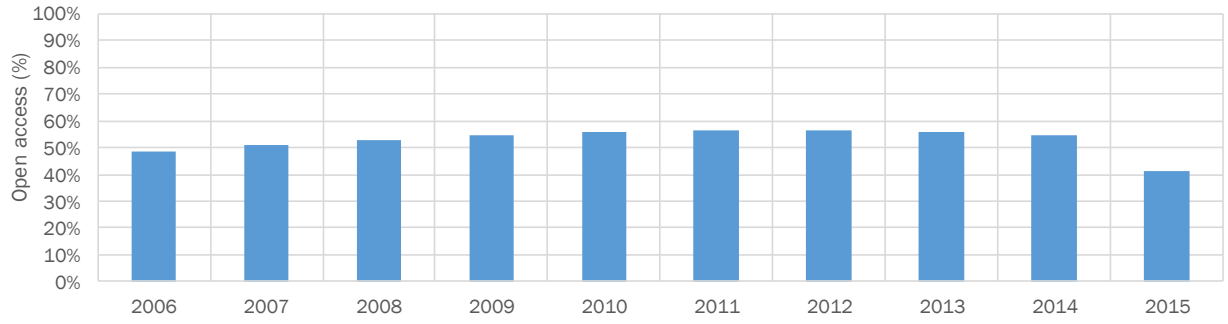


Figure 1 Percentage of OA per publication year (2006–2015), as measured in Q3 2016

Note: Data are presented according to publication year. Underlying data available in Table IX.
Source: Prepared by Science-Metrix using the Web of Science (Clarivate Analytics) and the 1science database

When observing the graph in Figure 1, it is important to realize that one is not seeing how OA looked at the time for all these years. It is a representation of the number of papers available for these different years at the time the measurement was taken (Q3 2016). For example, this measure suggests that about 50% of the papers published in 2006 can now be found for free, in an unencumbered manner. That said, it should be made clear that 10 years ago, if someone had looked at the proportion of papers available for the latest year, it would have been much lower.

Two factors account for this increase over time. First, the number of papers from 2006 available in OA has augmented as embargoes have ended. A second temporal effect is the backfilling of publications that continue to be added for earlier years. Older publications are added online each year, which contributes to increasing open access levels for earlier years. These patterns of populating previous years could eventually result in most publications older than a certain age being freely available online.

To characterize backfilling, it is necessary to have access to a large number of snapshots of publications' open access status. This requires making multiple assessments of the open access status of publications over long periods of time, as well as keeping records of the changes in status of each individual publication. Hence, to adequately study the growth of OA availability with time series, it would be necessary to use trends based on the production year (or date) of the snapshots performed in a longitudinal manner. Instruments should be in place to start such a series of rigorously produced snapshots by the time the *SEI 2018* edition is being prepared, though it won't be possible to go back in time to track changes that occurred before the first snapshots were taken.

As measured by the 1science database (applying a calibration factor of 1.2), the open access availability of scientific publications indexed in the WoS is currently about 50% for papers published in 2006 and reaches close to 60% for publication year 2011 (Figure 2).[24] As a result, for recent years, more than half of the scientific output indexed in the WoS can be retrieved online for free. Open access availability for the United States and China, both being by far the world leaders in terms of scientific output, is quite different.

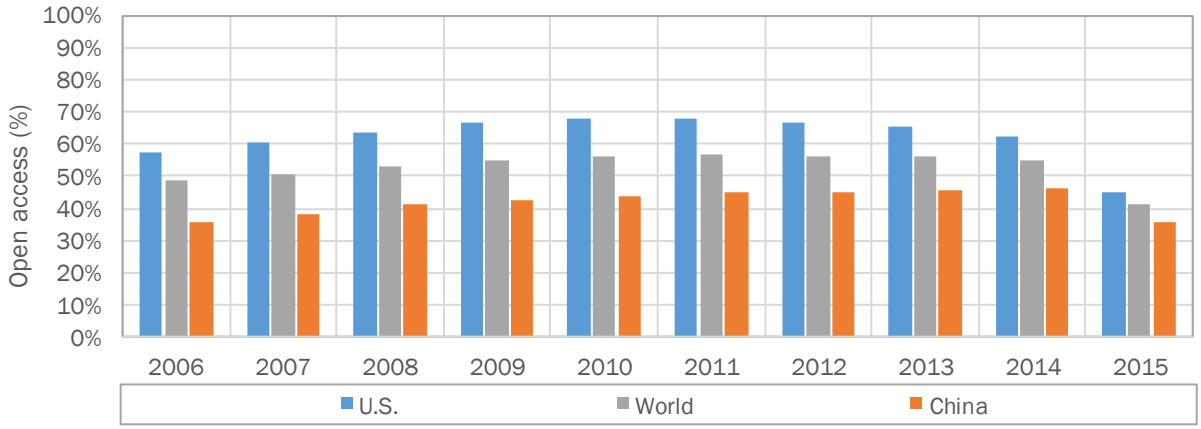


Figure 2 Percentage of OA per publication year (2006–2015), for the United States, China and the world, as measured in Q3 2016

Note: A 120% calibration factor was applied to the raw measures taken from the 1science database. Underlying data available in Table X.

Source: Prepared by Science-Metrix using the Web of Science (Clarivate Analytics) and the 1science database

About 70% of the papers published in 2010 and 2011 having at least one U.S. author are now available for free, whereas about 50% of the papers having at least one Chinese author and published in recent years are now freely available. For convenience, the same calibration factor of 1.2 has been used for papers from the United States and from China. However, as shown in the previous section of this report, a greater

[24] 2014 might still be missing some publications because of embargo periods or insufficient time for researchers to make these publications available online.

correction should be applied to China because of the likelihood of technical aspects being at play that have the effect of underestimating the OA availability of Chinese papers. Consequently, these measures are floor values in the case of China, but closer to reality in the case of the United States, albeit still somewhat conservative as the correction factor applied here is akin to a lowest common denominator, and most of the measures should be corrected somewhat more.

Brazil comes in first in OA availability among the countries with the largest number of papers indexed in the WoS, with three-quarters of its publications published between 2008 and 2014 being currently available, discoverable and free to download (Table IV). This is certainly helped by the SciELO repository, which comes in second as the main source of Brazilian open access articles (almost tied with ResearchGate, data not shown). The Netherlands also has about three-quarters of its papers available for free, whereas Switzerland has about 70% of its authored papers in OA. The United Kingdom, Sweden and the United States all have about two-thirds of their papers freely available. Countries that are lagging somewhat include Russia (45%) and China (46%). Other OA leaders generally have between half and 60% of their papers available for free.

Table IV Percentage of OA per publication year (2006–2015), per country, as measured in Q3 2016

Country	2006	2007	2008	2009	2010	2011	2012	2013	2014	2015
World	**48%**	**51%**	**53%**	**55%**	**56%**	**57%**	**56%**	**56%**	**55%**	**41%**
Brazil	64%	70%	73%	74%	76%	77%	77%	75%	74%	62%
Netherlands	67%	70%	72%	72%	74%	74%	73%	70%	68%	56%
Switzerland	59%	61%	64%	66%	67%	68%	68%	69%	67%	54%
United Kingdom	55%	58%	60%	62%	64%	64%	64%	65%	67%	57%
Sweden	51%	53%	56%	58%	62%	65%	64%	66%	66%	54%
France	53%	55%	58%	60%	63%	64%	64%	65%	64%	51%
United States	**57%**	**60%**	**64%**	**66%**	**68%**	**68%**	**67%**	**66%**	**63%**	**45%**
Italy	53%	55%	57%	58%	61%	62%	62%	63%	62%	48%
Poland	44%	46%	46%	48%	51%	55%	58%	60%	62%	48%
Spain	55%	56%	58%	59%	62%	63%	63%	63%	62%	48%
Australia	55%	56%	58%	60%	62%	61%	63%	62%	61%	48%
Canada	55%	57%	59%	60%	61%	61%	61%	61%	60%	45%
Germany	48%	50%	53%	54%	56%	57%	58%	58%	57%	46%
Turkey	40%	41%	45%	48%	51%	52%	50%	53%	54%	40%
Iran	41%	46%	50%	52%	56%	57%	55%	53%	51%	35%
Japan	44%	45%	48%	49%	51%	51%	52%	51%	50%	39%
Rep. of Korea	40%	47%	48%	50%	50%	50%	50%	50%	49%	38%
India	41%	41%	43%	47%	49%	49%	49%	49%	49%	35%
China	36%	38%	41%	43%	44%	45%	45%	46%	46%	35%
Russia	36%	37%	38%	39%	41%	41%	43%	45%	45%	33%

Source: Prepared by Science-Metrix using the Web of Science (Clarivate Analytics) and the 1science database

Outside this selection of leading countries based on scientific output, countries with lower levels of output often outperform others in terms of open accessibility (e.g., Gambia, Gabon, Kenya, Uganda, Papua New Guinea, data not presented). This finding is quite interesting, but it is difficult to determine its cause. One hypothesis is that researchers from these countries are perhaps less concerned about the reputation of the journals in which they publish and may be more disposed to publish in smaller, less-established gold open

access journals than their colleagues from other countries who focus on prestigious journals, which are more likely to be subscription-based journals that started publication a long time ago. A second hypothesis is that smaller countries also usually present higher levels of international collaboration, thereby increasing the chances of their publications being made available online by either themselves or one of their multiple collaborators, leading to higher chances of finding their publications online. A third hypothesis would be that these countries might be specializing in research topics where open access levels are higher.

3.2 Green and gold open access

As discussed earlier in this report, publications can be made available online following two methods: one coming from publishers themselves (i.e., gold open access) and the other from researchers using institutional or thematic repositories, personal websites and even academic social networks (i.e., green open access). Both approaches are used increasingly widely to diffuse scientific knowledge. Over the years, as the concept of open access has evolved, habits regarding open access also seem to have followed suit as mentalities changed and publishers have been forced to adapt to the disruptive effect of open access on their business model. These changes can be observed at Figure 3.

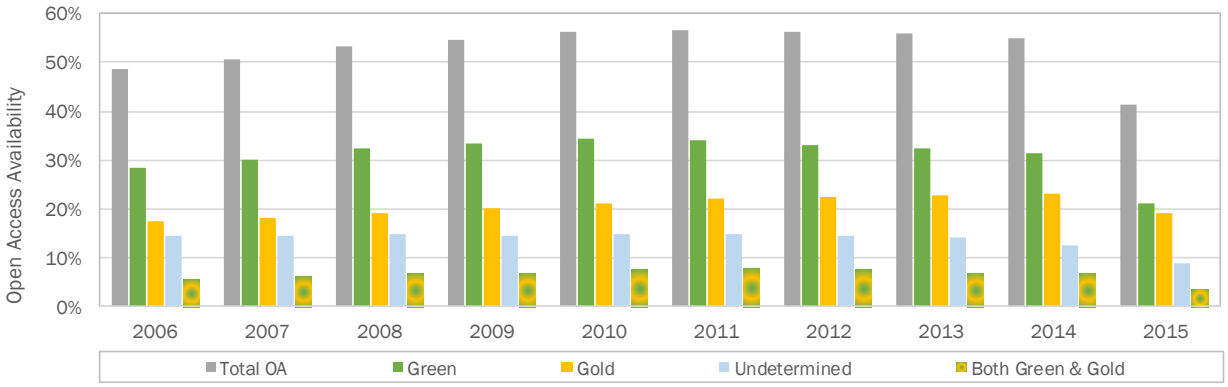


Figure 3 Percentage of OA per publication year (2006–2015), per OA type, as measured in Q3 2016

Note: Open access types are not mutually exclusive. Underlying data available in Table XI.
Source: Prepared by Science-Metrix using the Web of Science (Clarivate Analytics) and the 1science database

Using a recent snapshot of OA availability (Q3 2016), one can see availability is greater for green OA, and close to one-third of papers published in recent years have now been self-archived. As noted previously, ResearchGate is currently the largest venue for self-archiving. About one paper out of four is made available for free by the publishers themselves (gold OA), most of the time on their own websites but also frequently mediated by websites such as PubMedCentral, SciELO in some Romance-language countries, and JStage in Japan. About 8% of the gold OA papers are also self-archived by researchers or other parties such as librarians.

Because of the sizeable work involved in determining the types of open access for all sources of freely available papers, a portion of open access publications hasn't been assigned a type yet. About 10% of all publications in the 1science database are available in open access but their OA type is still unknown. This

share could impact the trends presented for green and gold open access if unknown sources are more heavily skewed toward one type or another.

As more sources are coded as either green or gold, numbers will become more robust. Figure 3 presents non-conclusive evidence that gold open access provided by publishers could be gaining ground and could potentially surpass green open access in the near future. However, there are several factors to consider before arriving at that conclusion. First, as mentioned before, the snapshot of OA availability shows the effect of the lag time required for papers to become available after an embargo period and for backfilling efforts to produce tangible results. Gold OA is far less likely to be affected by these factors than green OA, which suffers from embargoes imposed by publishers who want to maintain their revenues with paywalled journals.

Also, the recall of the 1science database is in the 60% to 80% range, and the distribution between green and gold OA of missing publications could be skewed toward either OA type. As already mentioned, discovering and harvesting green OA presents greater challenges than for papers in gold OA as the dispersion level is substantially greater for green. Hence, it is possible that green OA is more widely underestimated than gold.

Table V examines OA availability by type (with a conservative 1.2 calibration applied to the 1science database measures). Although gold and green open access account for similar shares of the total open access in health sciences (about 30%) and arts and humanities (9% for green and 7% for gold, respectively), there are large gaps between OA types in natural sciences (37% and 15%, for green and gold respectively), applied sciences (29% and 13%), and economic and social sciences (21% and 8%). Additionally, the percentage of open access for which the type is unknown also varies considerably across domains, ranging from 9% of all publications under arts and humanities being available in open access of an unknown type, to more than 20% for economic and social sciences.

Table V Percentage of OA across scientific domains for publication year 2014, per OA type, as measured in Q3 2016

	Total OA	Green	Gold	Both Green & Gold	Undetermined
WoS	**55%**	**31%**	**23%**	**7%**	**12%**
Health Sciences	59%	30%	33%	9%	10%
Natural Sciences	55%	37%	15%	5%	12%
Applied Sciences	47%	29%	13%	3%	12%
Economic & Social Sciences	44%	21%	8%	1%	21%
Arts & Humanities	24%	9%	7%	1%	9%

Note: Percentages are based on the total number of publications and not only open access publications.
Source: Prepared by Science-Metrix using the Web of Science (Clarivate Analytics) and the 1science database

Table VI presents OA levels by type for the top publishing countries in 2014. Here again, Brazil stands out due to the presence of SciELO, which is an extremely effective platform used to diffuse scholarly papers published in Brazil and in many other countries where Romance languages are spoken.

Overall, one can observe a generally lower level of green open access for Asian countries. China, Japan and the Republic of Korea all have about a quarter of their papers appearing in green OA, compared to 35% to 45% for most Western countries.

Gold OA is not used all that frequently in India and Russia. However, one must be careful in the interpretation of these data as there are many scholarly journals published in India that are suspected of using predatory practices. There are many such practices: one of the most common is to simply claim that a journal is peer reviewed when in fact it is not, and all articles are published provided authors pay the article processing charges demanded. More research is needed to help determine the extent of this usage pattern, which is not measured here as most of these journals are likely not included in the 1science database, Scopus and the WoS.

Table VI Open access levels, by OA type, for the top publishing countries (2014)

	Papers	OA Total	Green	Gold	Undetermined
World	**1,490,237**	**55%**	**31%**	**23%**	**12%**
United States	397,773	63%	38%	24%	14%
China	281,277	46%	23%	22%	8%
United Kingdom	111,666	67%	36%	28%	28%
Germany	104,695	57%	36%	24%	14%
Japan	78,193	50%	24%	27%	11%
France	72,648	64%	46%	22%	14%
Canada	65,918	60%	36%	25%	14%
Italy	65,005	62%	42%	23%	13%
India	58,439	49%	34%	16%	8%
Australia	58,118	61%	38%	23%	18%
Spain	57,530	62%	38%	22%	18%
Rep. of Korea	54,977	49%	25%	25%	10%
Brazil	41,315	74%	42%	41%	11%
Netherlands	38,902	68%	42%	28%	21%
Russia	30,915	45%	35%	10%	9%
Switzerland	28,764	67%	41%	28%	23%
Iran	27,815	51%	32%	19%	9%
Turkey	27,324	54%	30%	22%	14%
Sweden	25,896	66%	43%	29%	19%
Poland	25,314	62%	34%	29%	14%

Note: These categories are not mutually exclusive, e.g., a gold paper can also have a self-archived version (green OA). Color gradient is applied against the world level, with values above colored in green and those below in red, with stronger intensity of the gradient indicating a larger departure from world observed level.
Source: Prepared by Science-Metrix using the Web of Science (Clarivate Analytics) and the 1science database

3.3 Citation advantage of open access publications

It is a well-documented fact that open access leads to higher citation levels.[25] Although other underlying factors could explain parts of this effect, the generalized higher impact for open access publications across countries and disciplines represents a strong indication of the existence of a citation advantage[26] related to open access. Whereas previous studies were based on samples of various sizes, in the current report some small and some fairly large analyses could be performed on the whole WoS database. Impact analyses in this report are based on one indicator: the average of relative citations (ARC), which is a normalized indicator of scholarly or scientific impact. To calculate this, citations to each publication are counted and then normalized against the average level of citations of all publications from the same subfield, year and document type to obtain a relative citation (RC) score. The average of relative citations is then simply the average of all RCs related to a specific entity (e.g., country, discipline).[27]

Figure 4 presents a contemporary snapshot of the citation impact for open access and non-open access articles published during the last decade. This cannot be seen as a typical time series, but rather as a view of the cumulative citations given to these papers in or before Q3 2016.

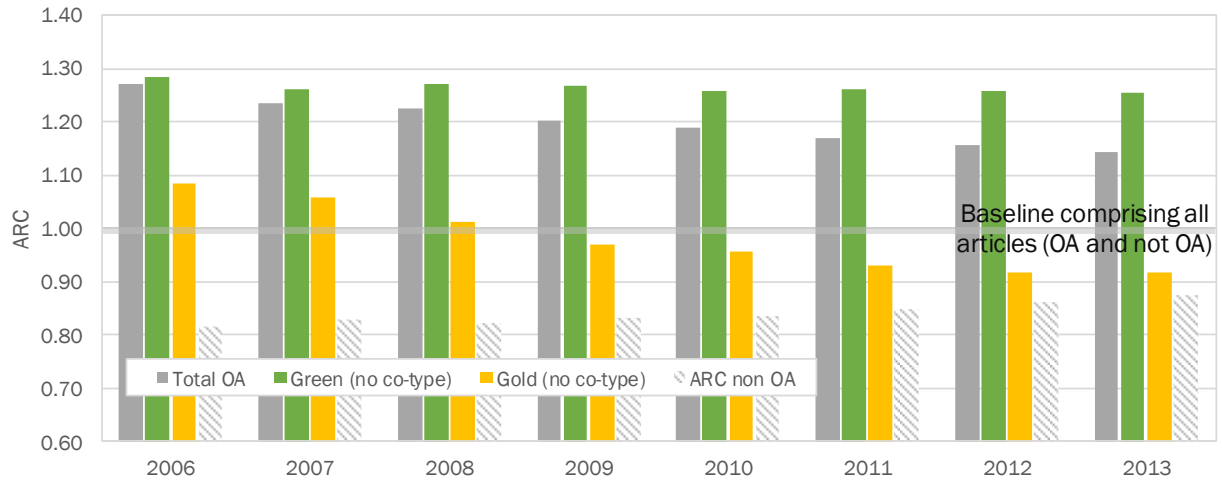


Figure 4 Scholarly impact (ARC), by OA type, for papers published between 2006 and 2013, as measured in Q3 2016

Note: Underlying data available in Table XII.
Source: Prepared by Science-Metrix using the Web of Science (Clarivate Analytics) and the 1science database

This figure highlights the citation advantage of open access articles, with open access always yielding a higher impact than non-OA. The interpretation of this figure is delicate because citations have not had time to cumulate as much for recent years, and we know that OA documents appear only progressively,

[25] See http://sparceurope.org/oaca_table/ and Archambault, E. et al. (2014). *Proportion of open access papers published in peer-reviewed journals at the European and world Levels—1996–2013*. Prepared for the European Commission by Science-Metrix. Retrieved from http://science-metrix.com/files/science-metrix/publications/d_1.8_sm_ec_dg-rtd_proportion_oa_1996-2013_v11p.pdf

[26] Citation advantage is calculated by dividing the average impact of open access publications by the average impact of non-OA publications.

[27] A minimum of three years of available citations are needed to compute a relative citation score; therefore, ARC scores for this study do not include publications published after 2013.

especially for green OA. Nonetheless, one can see a convergence of impact for OA and non-OA. This can be expected as the proportion of OA papers grows and these papers increasingly weigh on the average citation score. Although this needs to be taken for what it is, these data suggest a convergence of impact of OA and non-OA around 2024.

What is more difficult to interpret in Figure 4 is what appears to be a drop in the impact of gold OA. This figure does not tell the whole picture though. After data are added from an older measure performed for the European Commission by Science-Metrix in 2014, Figure 5 sheds light on a potentially complex phenomenon. One must be careful here in interpreting these data because the recent data for the NSF comprise only gold papers, which are not simultaneously available in green, whereas the data computed in 2014 for the European Commission comprised a proportion of papers having both gold and green (which as one can see in the data in Figure 3 represents about one-third of gold OA papers). The presence of a number of green OA papers in the 2014 data would have tended to increase the impact of these papers, so their presence is not of a nature to change what follows. In the last two years, gold OA papers have been increasingly cited compared to the whole population of papers of any type. For instance, the ARC of gold papers published in 2006 was 0.75, and it has now increased to 1.08. Likewise, papers published in 2011 and that were available in gold form had an ARC score of 0.58, whereas those 2011 papers available in gold in 2016 had reached 0.93.

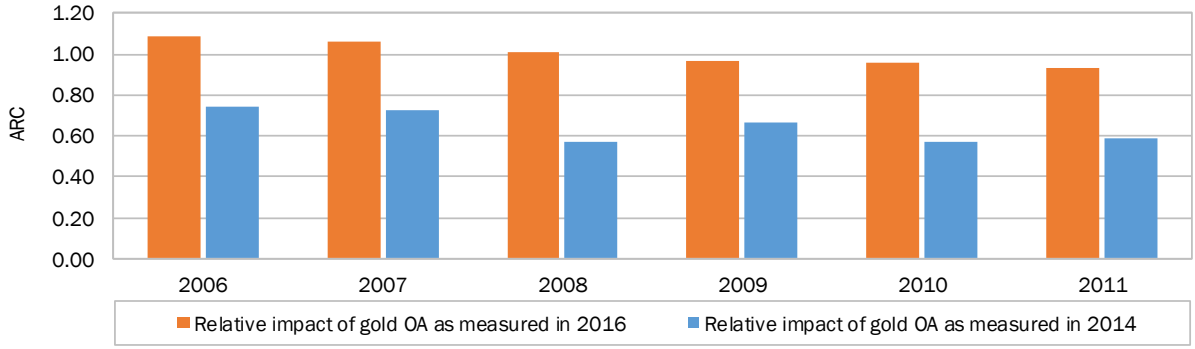


Figure 5 Scholarly impact (ARC) of gold OA for papers, published between 2006 and 2011, as measured in Q2 2014 and in Q3 2016

Note: Underlying data available in Table XIII.
Source: Prepared by Science-Metrix using the Web of Science (Clarivate Analytics) and the 1science database

These data suggest that more research is needed to fully understand how the citedness of different types of OA is evolving over time. There are many factors that could affect the observed score, including the changing proportion of different types of OA, the methods used to measure impact (using average scores may not be the best way to understand what is happening, and perhaps non-normalized data would be more telling), and finally, changes in the appreciation of different types of OA on the part of researchers. Stated in a different manner, the factors could be structural, methodological or social, or a combination of these. Understanding the impact of open access is crucial to policymaking and warrants more in-depth examination.

This citation advantage is present across all domains of scholarly communication, as displayed in Table VII. Although open access results in about 42% higher citation levels globally (citation advantage of 1.42),

with an effect ranging from 30% to 48% higher citation impact in applied sciences (citation advantage of 1.30), natural sciences (citation advantage of 1.39), and health sciences (citation advantage of 1.48), the effect is even stronger (75% and 99%) in the two domains related to social sciences and humanities (1.75 in economic and social sciences, 1.99 in arts and humanities). Furthermore, in most cases, green open access results in even higher levels of impact, particularly in arts and humanities (ARC = 2.00, citation advantage of 2.46) and applied sciences (ARC = 1.33, citation advantage of 1.51), and slightly higher in natural sciences (ARC = 1.26, citation advantage of 1.49). However, this is not the case in economic and social sciences (ARC = 1.29) and health sciences (ARC = 1.19), for which the impact of green open access is slightly below or on par with the impact level of all open access publications.

One important finding is that although, as shown in Figure 4, gold open access publications presented a slightly higher impact level than non-open access publications in recent years, data for 2010 tend to demonstrate that the impact of gold open access is pulled up by health sciences, as it is the only domain for which gold publications present higher impact than non-OA publications. Although health sciences' gold publications present an ARC score of 1.11, compared to 0.81 for health sciences' non-OA publications (citation advantage of 1.37), the other domains all present lower scores for gold publications.

Table VII Impact of open access publications, by OA type, at the level of scientific domains (2010)

Domain	All papers	All OA	Green (no co-type)	Gold (no co-type)	Non OA	Ratio ARC OA/Non OA
WOS	**1.00**	**1.19**	**1.26**	**0.96**	**0.84**	**1.42**
Applied Sciences	1.00	1.16	1.33	0.69	0.89	1.30
Arts & Humanities	1.00	1.62	2.00	0.67	0.82	1.99
Economic & Social Sciences	1.00	1.33	1.29	0.74	0.76	1.75
Health Sciences	1.00	1.19	1.19	1.11	0.81	1.48
Natural Sciences	1.00	1.17	1.26	0.80	0.85	1.39

Note: Citation advantage is the ratio of ARC OA/ARC non-OA. Color gradient is applied against the world level, with values above colored in green and those below in red, with stronger intensity of the gradient indicating a larger departure from world reference.
Source: Prepared by Science-Metrix using the Web of Science (Clarivate Analytics) and the 1science database

At the country level, the citation advantage is widespread for open access publications (Table VIII). For articles published in 2010, all countries among the top 20 publishing for that year present a citation advantage related to their open access publications. Although other factors could play a role in these higher levels of citations (such as international collaborations), the fact that all countries present the same effect is a strong indication that open access does indeed result in higher citations and that this is so across the board. The highest citation advantages across leading countries are observed for Russia (citation advantage of 2.65), Poland (citation advantage of 1.70) and France (citation advantage of 1.66).

In the future, it would be interesting to examine longitudinal data to understand how the OA level of different countries has contributed to changing their place in the scholarly communication system. For instance, has the higher level of OA contributed to increasing the relative ranking of Brazil and other countries that used SciELO compared to countries who made little effort to make their content openly available? Conversely, has the presence of so many gold journals contributed to a ghettoization of Brazilian

science (as seen through the large number of journals in SciELO), or of Japanese science (as seen through the large number of journals in JStage) for that matter?

Table VIII Impact of open access publications, by OA type, at the country level (2010)

	All papers	OA vs. non OA			Green (no co-type)		Gold (no co-type)	
		OA	Non OA	Ratio ARC OA/Non OA	Score	Ratio ARC Green/Non OA	Score	Ratio ARC Gold/Non OA
World	**1.00**	**1.19**	**0.84**	**1.42**	**1.26**	**1.51**	**0.96**	**1.14**
United States	1.32	1.53	1.04	1.46	1.52	1.46	1.39	1.33
China	0.99	1.18	0.88	1.34	1.35	1.54	0.81	0.93
United Kingdom	1.34	1.57	1.09	1.44	1.49	1.37	1.54	1.42
Germany	1.24	1.55	0.96	1.61	1.53	1.59	1.45	1.51
Japan	0.88	1.01	0.79	1.29	1.22	1.55	0.77	0.98
France	1.19	1.47	0.89	1.66	1.39	1.56	1.45	1.64
Canada	1.27	1.48	1.06	1.40	1.45	1.37	1.33	1.26
Italy	1.17	1.38	0.94	1.47	1.30	1.38	1.35	1.43
Spain	1.13	1.33	0.91	1.46	1.34	1.47	1.14	1.25
India	0.74	0.82	0.69	1.19	0.98	1.42	0.50	0.72
Rep. of Korea	0.90	1.01	0.82	1.22	1.26	1.53	0.69	0.83
Australia	1.31	1.52	1.08	1.40	1.45	1.34	1.48	1.37
Brazil	0.73	0.75	0.71	1.06	1.04	1.46	0.46	0.65
Netherlands	1.52	1.66	1.28	1.30	1.56	1.23	1.60	1.25
Russia	0.51	0.87	0.33	2.65	0.78	2.38	0.91	2.77
Turkey	0.66	0.72	0.61	1.18	0.90	1.48	0.45	0.74
Switzerland	1.55	1.80	1.22	1.47	1.72	1.40	1.72	1.41
Sweden	1.37	1.58	1.15	1.38	1.48	1.29	1.56	1.36
Poland	0.72	0.94	0.55	1.70	1.02	1.83	0.72	1.30
Belgium	1.39	1.60	1.11	1.44	1.47	1.32	1.81	1.63

Note: Citation advantage is the ratio of ARC OA/ARC non-OA. Color gradient is applied against the world level, with values above colored in green and those below in red, with stronger intensity of the gradient indicating a larger departure from world reference.

Source: Prepared by Science-Metrix using the Web of Science (Clarivate Analytics) and the 1science database

4 Conclusion

This report has provided the following operational definition of open access:

> In the present study, articles are considered as being gratis OA if they are available on the public Internet (i.e., sites that don't require a registration) in full-text form, and can be read and downloaded for free. One can also add that as the BOAI specifically mentioned "use them for any other lawful purpose," anonymous websites whose raison d'être and modus operandi are primarily to diffuse illegally obtained scientific literature are excluded from the present measurement.

This section briefly discusses the findings of the report, starting with the characterization of the data sources, then examining the preliminary measures obtained and what needs to be done to improve them, and finally coming back to the definition of open access used thus far.

4.1 Characterization of the 1science database using Scopus and WoS data

This report has detailed population-level measurements of the open access availability of publications indexed in two bibliometric databases—the Web of Science (WoS) by Clarivate Analytics and Scopus by Elsevier. This was achieved by matching the database populations to the 1science database to determine the availability of the papers in OA form.

In doing this, it was important to perform a careful characterization of the 1science database for two reasons. The first is that this database was originally designed for another use: that of creating a coherent collection of articles published in peer-reviewed or quality-controlled journals for which at least one version of the full text could be downloaded for free, in an unencumbered manner.

The second reason is that the 1science database uses a stricter definition of open access than that used in the present report. In the present report, documents are considered OA if they can be downloaded manually by a human, provided the documents are not on an anonymous site whose mission is to diffuse articles regardless of legal aspects (such as thirdworld.nl or Sci-Hub), and provided that they can be downloaded without registering with a site (such as would be the case on Facebook and sometimes on academic social networks). Obviously, large-scale discovery systems such as Google Scholar and the 1science database cannot be built through manual work, and automated processes require a stricter definition of open access, including that websites can be crawled. This characterization revealed the need to use calibrated measures to better align the results obtained with the measurement method to that of the gratis OA definition used in the present study.

A comparative analysis of the recall and precision levels of the 1science database was performed using Scopus and the WoS. This helped to characterize the 1science database and identify some of its current limitations. Two policy-relevant indicators were selected for in-depth analyses: country affiliation of authors on publications, and scientific disciplines. These indicators were selected because they are very frequently used in bibliometric studies, including those performed by the NSF, and they appear in the NSF's SEI.

Although the 1science database's precision (capacity to not include records that shouldn't be included, or avoiding false positives) was stable overall, recall (capacity to include all relevant records, or avoiding false negatives) varied substantially across country affiliation, scientific domains and the database selected as a baseline (i.e., the WoS or Scopus).

Characterizing the 1science database using WoS- and Scopus-indexed papers with authors from the United States and China showed that recall for Chinese publications was lower than for the United States in both databases, which resulted in a larger underestimation of open access papers from China using the 1science database. This phenomenon was particularly strong in the Scopus database and was somewhat less pronounced in the WoS. There are multiple reasons for this. Scopus covers more Chinese journals than the WoS, and 1science had not found an authoritative list of quality-controlled or peer-reviewed Chinese journals that could be used to whitelist these journals in the 1science system. This is important as, in contrast to Scopus and the WoS, which sometimes include non-academic or non-peer-reviewed journals such as *US News*, *The Economist* and *Scientific American*, 1science includes only scholarly journals that have undergone peer review or academic editorial control. Therefore, many journals in Scopus and the WoS have not yet been whitelisted by 1science, and some never will be.

There are also technical challenges linked with language and character sets. Journals that contain Chinese characters, which could have been more likely to have Chinese metadata (on Chinese sites), present a technical challenge to both the harvester and the data processing pipeline used to build the 1science database index. Several articles that were not seen by Science-Metrix' analysts could be in the 1science index but failed to be matched to Scopus and the WoS.

This exploratory analysis shows that the measurement protocol used in the present study encounters more challenges with papers published in languages other than English and using character sets other than Latin. This limitation is a general one as all mainstream bibliographic databases currently used in Western countries are optimized for searching Western-language articles, particularly those written in English, which is widely assumed to be the lingua franca of science. The widespread assumption that the most important work is published in English will be increasingly tested as more and more open access papers from non-Western countries become discoverable because of the growing availability of linguistic computing technology and the use of linguistic skills not frequently used in Western companies.

Examining different domains of scholarly activity reveals the challenges in representing the full extent of output in the arts and humanities, where only about 67% of the papers could be found with the present measurement protocol, compared to 76% to 82% for other domains. Lower recall for arts and humanities is, once again, the result of multiple factors. The main factor is that articles in arts and humanities are available on a larger variety of sites, many of them comprising only a few items. Once found, papers in arts and humanities sometimes appear on websites where metadata are not structured in a way that facilitates their automated aggregation, compounding the problem. As a consequence, arts and humanities articles require a substantially greater effort to be discovered and harvested, yet they constitute only a small proportion of the stock of scholarly papers. Moreover, because there are fewer authors on arts and humanities papers compared to articles from other domains, arts and humanities articles tend to be available in fewer places. When there are 10 authors on an article, more authors can be active in archiving the paper in different places compared to an article with a single author.

4.2 Open access measures

Preparing statistics on open access for countries and scientific disciplines will necessarily lead to underestimations of open access levels for the foreseeable future if proper calibration is not used. In Section

3 of this report, a conservative calibration factor of 1.2 was used across the board when computing proportions. This simple "translation" of availability scores was a "good enough" effort for a scoping report, but more sophisticated calibration techniques could be used for future measurement. In this respect, the next round of statistical production should, ideally, do the following:

1. Use a sizable random sample that can be used to determine calibration factors to apply in different cases: at the country level, at the field level, etc.
2. Determine whether and how to use multiple calibration factors when drilling down into data (e.g., combining calibration for domain-level and country-level analyses)

The data presented in the present report confirm that open access publishing is continuing to gain ground. Some countries have three-quarters of their papers that can be downloaded for free in an unencumbered manner on the Internet, and most of the world's leading countries in research have more than 50% of their papers available for free. For instance, at least two-thirds of U.S. authored papers published between 2010 and 2013 could be found in gratis open access form in 2016. In the case of Brazil, at least three-quarters of the papers from the same period were gratis OA in 2016.

There is no doubt that the world of scholarly communication is firmly engaged in the digital revolution and that it has passed the tipping point where more than 50% of articles are available in open access. Much is still unknown about open access, and more measurement and research are needed to understand the effects and functioning of OA mandates and policies. There is a need to take regular snapshots of OA availability to examine its evolution, as point-in-time measures can only provide for stationary analyses. There is also a need to measure in greater depth how mandates are being implemented. In particular, where are researchers archiving their papers? What is the respective contribution of institutional repositories, subject-based repositories, academic social networks, and publishers shifting from a strictly subscription-based model to one based on hybrid and gold OA? And how is the increasing amount of money spent on article processing charges for hybrid journals reflected in the cost of subscription to these journals by libraries?

Appendix — Data underlying report figures

Table IX Underlying data for Figure 1

Publication Year	Papers		Share	
	Total	Open access	Raw	Calibrated
2006	980,477	395,927	40.4%	48.5%
2007	1,050,083	443,927	42.3%	50.7%
2008	1,129,441	499,992	44.3%	53.1%
2009	1,183,706	539,306	45.6%	54.7%
2010	1,226,929	574,191	46.8%	56.2%
2011	1,308,110	616,958	47.2%	56.6%
2012	1,375,335	644,512	46.9%	56.2%
2013	1,451,327	676,835	46.6%	56.0%
2014	1,490,237	680,981	45.7%	54.8%
2015	1,455,361	502,158	34.5%	41.4%

Note: Data are presented according to publication year.
Source: Prepared by Science-Metrix using the Web of Science (Clarivate Analytics) and the 1science database

Table X Underlying data for Figure 2

Publication year	World	United States	China
2006	40.4%	47.8%	29.9%
2007	42.3%	50.3%	31.9%
2008	44.3%	53.1%	34.4%
2009	45.6%	55.4%	35.6%
2010	46.8%	56.4%	36.5%
2011	47.2%	56.4%	37.4%
2012	46.9%	55.7%	37.6%
2013	46.6%	54.8%	38.2%
2014	45.7%	52.1%	38.4%
2015	34.5%	37.6%	29.5%

Note: A 120% calibration factor was applied to the raw measures taken from the 1science database.
Source: Prepared by Science-Metrix using the Web of Science (Clarivate Analytics) and the 1science database

Table XI Underlying data for Figure 3

Year	% OA	%Green OA	% Gold OA	% Green-Gold OA	% Unknown OA
2006	48.5%	28.4%	17.4%	5.7%	14.6%
2007	50.7%	30.2%	18.3%	6.1%	14.7%
2008	53.1%	32.3%	19.3%	6.8%	14.8%
2009	54.7%	33.3%	20.2%	6.9%	14.5%
2010	56.2%	34.3%	21.2%	7.4%	14.8%
2011	56.6%	34.2%	22.1%	7.9%	14.8%
2012	56.2%	33.0%	22.4%	7.4%	14.5%
2013	56.0%	32.3%	22.6%	7.0%	14.3%
2014	54.8%	31.5%	23.3%	6.9%	12.4%
2015	41.4%	21.0%	19.3%	3.6%	9.1%

Note: Open access types are not mutually exclusive.
Source: Prepared by Science-Metrix using the Web of Science (Clarivate Analytics) and the 1science database

Table XII Underlying data for Figure 4

Year	ARC	ARC OA	ARC Green OA (no co-type)	ARC Gold OA (no co-type)	ARC non OA
2006	1.00	1.27	1.28	1.08	0.82
2007	1.00	1.24	1.26	1.06	0.83
2008	1.00	1.22	1.27	1.01	0.82
2009	1.00	1.20	1.27	0.97	0.83
2010	1.00	1.19	1.26	0.96	0.84
2011	1.00	1.17	1.26	0.93	0.85
2012	1.00	1.16	1.26	0.92	0.86
2013	1.00	1.14	1.26	0.92	0.88

Source: Prepared by Science-Metrix using the Web of Science (Clarivate Analytics) and the 1science database

Table XIII Underlying data for Figure 5

Publication year	ARC of Gold OA	
	Measured in 2014	Measured in 2016
2006	0.75	1.08
2007	0.73	1.06
2008	0.57	1.01
2009	0.66	0.97
2010	0.57	0.96
2011	0.58	0.93

Source: Prepared by Science-Metrix using the Web of Science (Clarivate Analytics) and the 1science database