THE TIPPING POINT – OPEN ACCESS COMES OF AGE

Éric Archambault

Eric.archambault@science-metrix.com
Science-Metrix Inc., Observatoire des Sciences et des Technologies (OST),
Centre Interuniversitaire de Recherche sur la Science et la Technologie (CIRST),
Université du Québec à Montréal, Montréal, Québec, Canada

Abstract
The Open Access (OA) model for scientific publications has been examined for years by academics who have argued that it presents advantages in increasing accessibility and, consequently, in increasing the impact of papers. It has been noted that OA availability has increased steadily over the years. However, current measurement has seriously underestimated the proportion of OA peer-reviewed articles. This paper presents the results of a pilot study that shows evidence that the proportion of measured OA is so close to 50% that we have most likely passed the tipping point, that is, the stage where the majority of articles become available for free.

Conference Topic
Topic 10**:** Open Access and Scientometrics

Introduction
Interest in the academic community for Open Access (OA) publications has been increasing. The initial interest in the use of bibliometric methods focused on accessing the so-called citation advantage of OA as opposed to subscription-based journals (Antelman, 2004; Harnad & Brody, 2004; Craig, 2007). The literature of the time recognised a clear citation advantage to papers available in OA as opposed to papers diffused solely through subscription-based journals. Strong advocacy by authors such as Harnad (2003, 2008, 2012) suggested that benefits would ensue from so-called green OA, that is, research papers self-archived by their authors in various types of repositories. Unsurprisingly, in this context, librarians and information scientists noted that they had a new mission, which meant setting up and curating OA repositories (Proser, 2003; Bailey, 2005; Chan, Kwok, & Yip, 2005; Chan, Devakos & Mircea, 2005 Repanovici, 2012).
A part of the OA literature has discussed how authors and researchers (Pelizzari, 2004; Swan & Brown, 2004; Dubini, Galimberti & Micheli, 2010) and publishers (Morris, 2003; Regazzi, 2004) would react to this new paradigm. Evidently, business and economic models were discussed (Bilder, 2003; Kurek, Geurts & Roosendaal, 2006; Houghton, 2010; Lakshmi Poorna, Mymoon & Hariharan, 2012), but there was also interest in what models academia and libraries would follow (Rowland et al., 2004; Swan et al., 2005; Hu, Zhang & Chen, 2010).

As OA continued to make inroads, a growing number of papers examined the state of development of OA in specific countries (Nyambi & Maynard, 2012; Sawant, 2012; Woutersen-Windhouwer, 2012; Miguel et al., 2013) and in specific fields of research (Abad-Garcí et al., 2010; Gentil-Beccot, Mele, & Brook, 2010; Charles, & Booth, 2011; Henderson, 2013). In this context, it was not surprising to find papers that addressed the general question of OA availability as a proportion of the scientific literature, and the proportion of OA papers available in different fields of science (Björk et al. 2010; Gargouri et al., 2012).
This paper re-assesses OA availability in 2008 through a careful examination of recall, which leads to a doubling of the proportion of OA estimated by Björk et al. and by Gargouri et al. The paper argues that the tipping point for OA has been reached and that one can expect that, from the late 2000s onwards, the majority of published academic peer-reviewed journal articles were available for free to end-users. The paper presents data for 22 fields of science as well as for the European Research Area countries, Brazil, Canada, Japan, and the US.

Methods

Accuracy and Precision: The paper presents the results for the pilot phase of a study that aims to estimate the *proportion of peer-reviewed journal articles which are freely available, that is, OA* for the last ten years (the pilot study is on OA availability in 2008). It builds on two important concepts: (1) *accuracy*, reflected in the quality of the instruments used and the care taken in making measurements; (2) *precision*, which involves repeated measures, sampling and statistical analysis (see figure 1)—the later concept will be called *statistical precision* for reasons that will become obvious.

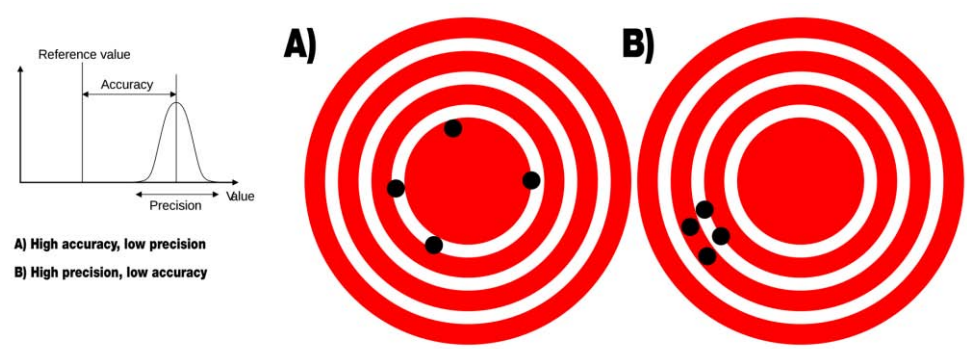

Figure 1. Accuracy and statistical precision (Adapted from http://en.wikipedia.org/wiki/Accuracy_and_precision)

Statistical precision can be approximated with the margin of error (ME). For a proportion (p) where the population is finite and known (N), is not systematically much larger than the sample size (n), and in which the values are discrete (for

example, papers), given a critical score Z (which will be set at 0.95 in the study), ME is calculated as follows:

$$ME = Z\sqrt{\frac{p(1-p)(N-n)}{n(N-1)}} + \frac{0.5}{n}$$

What complicates the use of these definitions is the need to examine accuracy with two more concepts used in information retrieval: recall and precision (hence the need to call the previous concept 'statistical precision'; the second precision-related concept will be knonw as 'retrieval precision'). Recall is the proportion of relevant records that are retrieved, while retrieval precision is the proportion of retrieved records that are relevant. If an instrument retrieves 25 records of which only 20 are relevant, and fails to retrieve 30 additional relevant records, its retrieval precision is 20/25 = 80% while its recall is 25/50 = 50%. Precision is synonymous with Type I errors (false positives), and recall with type II errors (false negatives). Thus, a high recall means that an instrument returned most of the relevant results, while high retrieval precision means that it retrieved more relevant results than irrelevant ones. Note that assessing the real positives accurately is frequently a distinct problem, as is the case in the present study.

Let us call π the proportion of the whole population of peer-reviewed papers that are OA. One cannot easily measure π directly because the population of scientific papers is relatively large, and there is currently no satisfactory complete repertory of that population. Hence, it is unlikely in the short term that someone will find another way than sampling to calculate p, an approximation of π. Though it is nearly impossible for p to equal π, it is the aim of this study to offer a robust design that will ensure that p is reasonably close to π. In the present study, two principal proportions will be calculated: (1) the overall proportion of OA literature; and (2) the proportion of the scientific literature published in gold journals. Before entering into the methodological details associated with the measurement as such, it is important to produce operational definitions of OA, green OA, gold OA and hybrid OA.

Types of OA scientific literature: Peter Suber suggests that '[o]pen-access (OA) literature is digital, online, free of charge, and free of most copyright and licensing restrictions.'[164] An effective definition of OA for this study is the following: 'OA, whether Green or Gold, is about giving people free access to peer-reviewed research journal articles.'[165] The following operational definitions of gold and green OA will be used in the present study.

- Gold OA refers to papers published in journals that provide free access to [peer-reviewed scholarly] papers. Authors sometimes, but not always, pay a fee for these publications. In the present study, Gold journals are those that provide cover-to-cover, instant access to articles.

[164] http://www.earlham.edu/~peters/fos/overview.htm.

[165] http://scholarlykitchen.sspnet.org/2011/09/07/oa-rhetoric-economics-and-the-definition-of-research/.

- Green OA generally refers to authors' self-archiving [of papers accepted in academic journals following a successful peer-review process].
- Hybrid OA is an increasingly important trend in scientific publishing by which authors pay for their papers to be available in OA in an otherwise not OA journal—'[h]ybrid open access journals provide Gold OA only for those individual articles for which their authors (or their author's institution or funder) pay an OA publishing fee.'[166]

A note on the concept of open-access versus toll-access literature is in order here. OA is rarely free, and can generally be seen as moving the toll plaza before the publication process as opposed to placing it after it. Open access will rarely entirely miss exacting a toll somewhere, be it on taxpayers' or on philanthropists' funds, or on the time of volunteers. Thus, the term toll-access, to distinguish the non-OA literature, is avoided here.

Peer-reviewed journal articles and original contributions to knowledge: A central part of the scientific literature is comprised of papers published in peer-reviewed journals (Larivière et al., 2006). This study concentrates on peer-reviewed, scholarly articles and omits the many other types of vehicles that are used for the *written* diffusion of scientific knowledge, namely books and conference proceedings, as well as research reports, mimeos and other heterogeneous forms, collectively called grey literature. A best practice in bibliometrics is to use only articles that can be considered original contributions to knowledge. The tradition in the Web of Knowledge (and its predecessor, the Science Citation Index) was to restrict the selection of document types to articles, notes and reviews (Carpenter & Narin, 1980). In Scopus, the tagging of articles is substantially more complex, and a combination of Source Type and Document Type is required to keep only what can be considered original contributions to knowledge. The present study uses the following operational definition: *articles that use references and are cited*. This definition, and empirically obtained thresholds, can be used to prune the Scopus production database of trade journals and non-original contributions to knowledge (at the macro level rather than at the article level to prevent the exclusion of papers that have not yet been cited). The resulting types of documents used are presented in the accompanying side box.

Source Type	Document Type
Book Series	Article
	Conference Paper
	Review
	Short Survey
Conference Proceeding	Article
	Review
Journal	Article
	Conference Paper
	Review
	Short Survey

Calculating the denominator: An important aspect of the project involves determining the proportion of OA papers by precisely estimating the number of

[166] http://en.wikipedia.org/wiki/Open_access.

OA peer-reviewed papers (the numerator) and dividing this by a carefully designed estimate of the number of peer-reviewed articles (the denominator) for each of the selected 22 disciplines and for the total literature. A decision was made to use the Ulrich periodical database to provide an estimate of the denominator, and these data and rights to use them were acquired for this study. The strengths and weaknesses of Ulrich data are well known: for example, some journals that should be classified as peer-reviewed are not (and the reverse is also true). A good example of this is the OA journal *Activités*, which mentions that 'Texts that have been submitted to *Activités* (www.activites.org/) will be assessed by two referees (called upon in view of the article). Each will give his or her opinion on the text.'[167] Despite this, and a description that clearly suggests scholarly content and the presence in papers of references to scholarly work, Ulrich has not classified this journal as refereed. Although several journals are likely to be classified 'Academic/Scholarly' in Ulrich and might be considered as contributing to science, this category cannot be included *en masse* as it comprises a substantial amount of material published in universities that has little scientific content. This is the case, for example, with the 'The Hilltop', classified by Ulrich as Academic/Scholarly, and claiming to be the 'The Student Voice of Howard University' (see http://www.thehilltoponline.com/). Consequently, the selection was restricted to Ulrich listed *journals* considered *refereed/peer-reviewed* AND Academic/Scholarly. Although imperfect, Ulrich remains the most extensive and authoritative and probably the least biased source of data on academic peer-reviewed journals and is therefore a solid calibration instrument for a systematic investigation of the peer-reviewed literature.

The core use of Ulrich in this project was to calibrate the proportion of papers from each of 22 disciplines used to present disaggregated statistics. The reason Ulrich is preferred is because *article-level* database publishers such as Elsevier (publisher of Scopus) and Thomson (Web of Science) are faced with choices having important commercial and profitability impacts. When selecting journals to be included for an article-level database such as Scopus, deciding whether to include a journal has a direct impact on production costs and partly because of this, database publishers tend to have a bias towards larger journals (economies of scale) and larger publishers (lowest transaction costs and economies of scale). However, whether a journal is small or large in terms of number of articles has substantially fewer consequences when it is included in a journal title database, where journal size can be expected to little impact on cost (some differences remain as it is likely easier to find information about the larger journals).

Ulrich cannot be used alone as it does not contain article-level information. The core work of the present project involved using a fully-licensed version of Elsevier's Scopus database hosted in house and conditioned over several years to produce bibliometric statistics. This meant that it was possible to randomly select papers among the millions of papers indexed. Ulrich was used to 'calibrate' the

[167] http://www.activites.org/resources/activites.eng.book.pdf.

proportion of peer-reviewed journal articles for each of the 22 fields used in this paper to present detailed statistics. The technique used to determine this proportion involved the following steps: 1) journals in Ulrich were matched to those contained in Scopus; 2) journals that intersected were given the discipline that was already contained in our classification of Scopus journals (for those that did not intersect, the Ulrich classification was compared with that used in our classification, and a matching table was used to attribute one of 22 disciplines to each of the journals); and 3) the number of articles per discipline was counted in the intersecting set, while the number of articles in the Ulrich set with no Scopus counterparts was determined by projecting the average number of articles for the 50% journals in Scopus with the fewest articles per journal. The reason for using the average number of articles for the 50% smaller journals is that experience has revealed that databases such as Scopus and the Web of Science index the largest journals first. For instance, the Web of Science covers about 12,000 journals, and Scopus about 18,000. Despite a 50% increase in journal coverage, Scopus only has about 20% more articles. A sensitivity analysis was performed to see the effect of calculating the average for the 75%, 50%, and 25% smallest journals (ranked by decreasing number of articles), and the results were broadly similar.

Strategy to measure the proportion of gold OA: Somewhat distinct strategies were used to calculate the occurrence of gold OA and total OA. For gold articles, an estimate of the proportion of papers was made from the random sample by matching the journals that were known to be gold in 2008. These journals were obtained from the Directory of Open Access Journals (DOAJ) and the list of OA journals in PubMed Central. This was done by matching journals' ISSN, E-ISSN and names from Scopus to the relevant records in the sample (the matching had about 100% precision, but recall may have been imperfect, hence the figures presented here can be considered a floor, rather than a ceiling).

Strategy to measure the OA proportion of scientific articles: Two samples and a sub-sample were produced to undertake a pilot study to measure OA availability in 2008 given the definition and assumptions presented above. A first sample of 20,000 was produced for early testing, and a sub-sample of 500 records was drawn from this sample to determine the availability of papers in OA using various search engines; a 'ground truth' was established by combining the validated results of these tests.
A second random sample of 20,000 records was drawn from Scopus and used to perform the measuring stage of the pilot study. This sample was restricted to papers published in 2008, and the results were restricted to original contributions to knowledge; records where the journal name or the record type contained a conference were excluded. Records for which the discipline was unknown were also set aside. The eligible record set from 2008, comprising somewhat more than 1.36 million records in Scopus, was 'tossed' five times using a pseudo-random method (using the newid() command in SQL Server), a subset of 100,000 records

was selected, placed in a subset, and tossed again. These 100,000 records were then imported into Excel, where a straightforward analysis of the distribution of the records by discipline was performed. This analysis showed that a subsample of 20,000 records would keep few records in three of the smaller disciplines (Philosophy & Theology, Visual & Performing Arts, and General Arts, Humanities & Social Sciences). For these disciplines, a random sample of 100 records was selected, and for the Built Environment & Design discipline, the 101 records that were part of the 100,000 records were all selected. As the objective was to produce a record set of 20,000, a subsequent selection was done for 19,599 records. These were selected by tossing the 100,000 a few more times using the rand() command in Excel, then proceeding to the selection of the required number of records.

Technique used to harvest OA articles: Although the pilot study was meant to build on the method pioneered by Björk *et al*. and human judgment was to be used in searching for and categorising the presence of OA, the pilot study has led to a gradual, but fundamental, modification of the original approach. After nearly two months of work, it became apparent that using professionals would be cost-prohibitive and too slow for a large scale study over several years.

A test was then conducted with 20,000 records being provided to the Steven Harnad team in Montreal. This relatively blind test produced recall that was good; the scores computed were much higher than those presented in previous papers, including results by Harnad's team. This was due to the use of Scopus, as opposed to the Web of Science as Harnad's team had done before. Some 500 records of this set were then extracted randomly and extensive testing was performed. The records were all searched manually in Google Scholar, Google, and Microsoft Academics. Records that could be downloaded for free and that came from any of these sources were considered OA, and the carefully verified sample a 'ground truth.'

These tests led to the following observations: Google Scholar and Google have substantial overlap, but each search engine has a somewhat distinct set of positive results. Microsoft Academics does not add much to the combined results of Google and Google Scholar. Importantly also, the results obtained suggest that the accuracy of the harvesting instrument, and the coverage of the database, are more important than a large sample size (statistical precision). For instance, the team led by Harnad measured only 22% of OA in 2008 overall 'out of the 12,500 journals indexed by Thomson Reuters using a robot that trawled the Web for OA full-texts' (Gargouri et al., 2012). Likewise, Bj rk *et al*. found a score of 20% using Scopus and Google as a search engine. When the Harnad team ran their robot on our Scopus sample, the proportion of total OA jumped to close to 32%, compared with the 22% they obtained in WoS as mentioned in their paper (this original sample was prepared rapidly for testing and might not have been perfectly random, so these results should be seen as tentative). This shows that a

technique to measure the proportion of OA literature based on the Web of Science produces fairly low recall and seriously underestimates OA availability.
Extensive testing was done with the subsample of 500 records. Because the original was not necessarily 100% random, this subsample cannot necessarily be considered as totally representative, but the results are nonetheless instructive. The results for the Harnad robot are as is and contain a few false positives, so the real positive score is actually lower. The Scholar, Google and Ground Truth results were manually validated and the documents downloaded, and as such, they can be considered accurate. The Ground Truth comprises the combined validated results from Google and Google Scholar in addition to one result from Microsoft Academics. Results from Microsoft Academics are not shown, as only the negative results from Scholar and Google were tested to examine whether this added any substantial results to the previous ones.

Table 1. Availability of OA in a sample of 500 Scopus records, 2008

Result	UQAM (Harnad)	Scholar	Google	Ground Truth
FALSE	350	293	290	262
TRUE	150	207	210	238
Total	**500**	**500**	**500**	**500**
% OA	30%	41%	42%	48%

Source: Computed by Science-Metrix

This extensive analysis therefore suggests that 48% of the literature published in 2008 may be available for free. Despite their high level of performance, neither Google nor Google Scholar can be expected to crawl the Web perfectly or to have a search engine so robust that it systematically presents all the relevant records in the first page of results (which we limited our analysis to), and hence cannot be expected to have a 100% recall, especially for academic articles (Arlitsch & O'Brien, 2012). Consequently, one can infer that OA availability very likely passed the tipping point in 2008 (or earlier) and that the majority of peer-reviewed/scholarly papers published in journals in that year are now available for free in one form or another to end-users.
These results suggest that using Scopus and an improved harvester 'to trawl the Web for OA full-texts' could yield substantially more accurate results than the methods used by Björk *et al*. and Harnad *et al*.

Results

Table 1 presents data on OA availability overall and for Gold journals (pure Gold, in that it does not include journals with an embargo period or traditional-model journals offering pay-per-article OA). Pay-per-article OA, journals with embargo periods and journals allowing partial indexing following granting agencies' OA policies are considered hybrid, and these data are bundled here with green OA (self-archiving). Papers in each of the 22 fields have been recalibrated given the method presented before (calibration based on Ulrich).The overall rate calculated

with the current harvesting instrument is 42% (plus or minus three percentage points). Considering that the instrument used has imperfect recall and considering that OA Gold journals are likely to be under-represented in both Scopus and the Ulrich database, this can be considered a floor rather than an upper limit.
OA availability varies considerably among disciplines. It seems that the tipping point has been passed (OA availability over 50%) in Biology, Biomedical Research, Mathematics & Statistics, and General Science & Technology. According to these data, a third or less of the papers can be found in OA in Chemistry, Enabling & Strategic Technologies, Historical Studies, and Engineering, while less than one paper out of five can be accessed free in Communication & Textual Studies and in Visual & Performing Arts. However, one must be careful with these last two figures as the statistical error is of the same order of magnitude as the measured proportion.

Table 2.Proportion of OA per discipline, 2008

Field	Papers	Green & Hybrid		Gold		OA	
		Papers	%	Papers	%	Papers	%
Agriculture, Fisheries & Forestry	780	199	26 ± 6	125	16 ± 7	324	42 ± 4
Biology	1,031	477	46 ± 4	161	16 ± 6	638	62 ± 3
Biomedical Research	1,618	858	53 ± 2	141	9 ± 5	999	62 ± 2
Built Environment & Design	100	30	30 ± 15	7	7 ± 25	37	37 ± 14
Chemistry	1,621	379	23 ± 4	154	9 ± 5	532	33 ± 3
Clinical Medicine	5,157	1,609	31 ± 2	501	10 ± 2	2,110	41 ± 2
Communication & Textual Studies	249	33	13 ± 19	15	6 ± 24	48	19 ± 17
Earth & Environmental Sciences	599	228	38 ± 5	28	5 ± 9	256	43 ± 5
Economics & Business	627	246	39 ± 6	23	4 ± 11	269	43 ± 5
Enabling & Strategic Technologies	1,267	301	24 ± 4	75	6 ± 5	376	30 ± 4
Engineering	1,168	290	25 ± 4	17	1 ± 8	307	26 ± 4
General Arts, Humanities & Social Sciences	25	11	44 ± 12	0.2	1 ± 70	11	45 ± 12
General Science & Technology	165	52	32 ± 9	40	24 ± 10	92	56 ± 6
Historical Studies	232	48	21 ± 13	20	9 ± 17	68	29 ± 12
Information & Communication Technologies	590	220	37 ± 5	30	5 ± 9	250	42 ± 5
Mathematics & Statistics	625	333	53 ± 4	31	5 ± 10	364	58 ± 4
Philosophy & Theology	164	52	32 ± 15	10	6 ± 27	62	38 ± 14
Physics & Astronomy	1,872	747	40 ± 3	89	5 ± 5	836	45 ± 3
Psychology & Cognitive Sciences	436	193	44 ± 6	17	4 ± 13	210	48 ± 6
Public Health & Health Services	581	194	33 ± 6	70	12 ± 8	264	45 ± 5
Social Sciences	1,051	313	30 ± 6	96	9 ± 8	408	39 ± 5
Visual & Performing Arts	43	7	16 ± 20	0.9	2 ± 44	8	18 ± 19
All Publications	**20,000**	**6,818**	**34 ± 4**	**1,649**	**8 ± 6**	**8,467**	**42 ± 3**

It is more delicate to interpret the proportion of Gold OA because of the large statistical error (resulting from the small sample and low occurrence). The overall Gold OA availability measured here is 8%, and this is generally consistent with the literature. Note however that this report uses a strict definition of Gold OA, and that many previous studies might have included disembargoed papers and pay-per-article OA, which is not the case here. Gold OA is widespread in General Science & Technology, Agriculture, Fisheries & Forestry, Biology, Public Health & Health Services, and Clinical Medicine. Less than 2% of the papers are

available in Gold journals in Visual & Performing Arts, Engineering and General Arts, Humanities & Social Sciences.

The prevalence of papers in hybrid forms (non-Gold) is especially high in Mathematics & Statistics and Biomedical Research. Less than one paper out of four can be found in hybrid forms in Engineering, Enabling & Strategic Technologies, Chemistry, Historical Studies and the Visual & Performing Arts.

A question that has animated OA advocates has been the so-called citation advantage of OA. This question is examined briefly in Table 3 using the Average of Relative Citation (ARC), a measure that reflects citation rates and is normalised to account for differences among scientific specialities in the propensity to use references and receive citations. These data present the relative citation rate of OA publications overall, Gold OA and hybrid OA forms relative to publications in each discipline. A score above 1 denotes that papers are more cited than in the field overall, while a score below 1 means that these publications are less frequently cited. For instance, papers in Agriculture, Fisheries & Forestry receive roughly the same level of citation (0.98) in OA overall than they do usually (the base measure is 1.0 for whole set of papers in a discipline). Importantly though, Gold OA papers are cited only half as frequently on average (0.49), although self-archived and other hybrid forms are cited 28% more frequently than the discipline's average (1.28).

Table 3. Scientific impact (ARC) of OA publications, 2008

Field	All Publications	Green & Hybrid	Gold	OA
Agriculture, Fisheries & Forestry	1.00	1.28	0.49	0.98
Biology	1.00	1.35	0.55	1.15
Biomedical Research	1.00	1.17	0.84	1.13
Built Environment & Design	1.00	1.07	0.25	0.91
Chemistry	1.00	1.18	0.38	0.95
Clinical Medicine	1.00	1.66	0.59	1.40
Communication & Textual Studies	1.00	1.23	1.55	1.33
Earth & Environmental Sciences	1.00	1.04	1.19	1.05
Economics & Business	1.00	1.39	0.07	1.28
Enabling & Strategic Technologies	1.00	1.37	0.64	1.23
Engineering	1.00	1.49	0.13	1.41
General Arts, Humanities & Social Sciences	1.00	1.28	0.00	1.25
General Science & Technology	1.00	2.60	0.40	1.64
Historical Studies	1.00	1.10	0.22	0.84
Information & Communication Technologies	1.00	1.50	0.73	1.40
Mathematics & Statistics	1.00	1.11	0.71	1.07
Philosophy & Theology	1.00	1.28	0.61	1.18
Physics & Astronomy	1.00	1.21	1.05	1.19
Psychology & Cognitive Sciences	1.00	1.21	0.86	1.18
Public Health & Health Services	1.00	1.31	0.68	1.14
Social Sciences	1.00	1.38	0.52	1.18
Visual & Performing Arts	1.00	1.15	n.c.	1.02
All Publications	**1.00**	**1.36**	**0.59**	**1.21**

An overall OA advantage occurs in all but four disciplines (Agriculture, Fisheries & Forestry; Chemistry; Built Environment & Design; Historical Studies). Gold OA only presents a citation advantage in three disciplines (Communication & Textual Studies; Earth & Environmental Sciences; Physics & Astronomy), and in those disciplines, except for one (Physics & Astronomy), the citation advantage is greater in Gold OA than in hybrid OA forms. Hybrid OA forms always present a citation advantage.

These data require careful interpretation. First, many Gold journals are younger and smaller, and these factors have an adverse effect on the citation rate and the ARC. Authors frequently prefer reading and citing more established journals, and it is a difficult endeavour to start a journal from scratch. It takes time to build a reputation and to attract established authors. It is possible though that Gold journals might provide an avenue for less mainstream, more revolutionary science. If so, the signature would be a much greater level of variation between the more highly cited papers and the baseline with no citation. Also, the ARC is not scale-invariant, and larger journals have an advantage as this measure is not corrected sufficiently for journal size (namely, it is not a scale-independent measure). So it might not always be the Gold nature of journals that lowers their 'citedness'; instead several structural aspects might be at play. Even so, the Gold journal industry is young, and it is still difficult to separate the wheat from the chaff. In this respect, it might be useful for authors to examine Beall's List of 'potential, possible, or probable predatory scholarly open-access publishers' to lower one's risk of spending money on journals that do not espouse scientific publishing best practices.[168]

A last aspect of the analysis based on the pilot study data is the examination of OA availability per country (for EU27, EFTA, Accession countries, ERA, and four comparables). Please note that fractional counting was used here as it was deemed as potentially providing a more precise portrait of the situation. In fractional counting, if two authors are from separate countries, each country is given half a publication. In contrast, full paper counting would have ascribed one paper to each country. One advantage of fractional counting is that one can add the fractions for all countries' output in a table and obtain a total. A drawback is that statistics might not seem as intuitive. In the table, the fractions of papers are presented only for scores below 10 (for example, 11 papers; 3.2 papers). The EU27, EFTA, and ERA all have roughly the same level of OA as that observed at the world level, though there are noticeable differences among countries.

Excluding countries with less than 50 papers (sum of all the fractions), the EU countries with the greatest OA proportions are the Netherlands, Finland, Romania, Portugal, and the United Kingdom. The countries with the lowest rate of OA adoption are Hungary, the Czech Republic, Poland, Germany, and Denmark. In countries outside the EU27, it is noteworthy that the US seems to have passed the tipping point (50%). Even more salient is the proportion of 62%

[168] http://scholarlyoa.com/publishers/.

observed in Brazil. This is no doubt due largely to the exemplary work performed by Scielo, which plays a key role in the Southern hemisphere in making scientific knowledge more widely available.

Discussion

One has to be careful when interpreting the results presented in this paper as the methodological instruments are not fully developed, and results could vary with growing accuracy. As a general rule, further development of the 'trawler' will increase recall and therefore, the proportion of OA presented here will surely increase. Sample size can be fine-tuned to obtain a satisfactory level of *statistical precision* as the margins of error presented above were certainly high in several areas. Future exercise will balance the sample more carefully to augment the number of papers from the smaller countries and the presence of papers from the smaller disciplines. We also endeavour to develop a robust method to distinguish more clearly between Gold OA, Hybrid OA and non-fully Gold journals, and self-archiving ('Green OA'). This presents many challenges, and statistics should be presented on the condition that they must not be too inaccurate. Other authors have presented results suggesting that OA availability was only half as high as carefully and prudently measured here, and this is certainly a reminder that it might be preferable to be reflective. Previous authors have measured what was in databases, or what search engines were able to do. Our goal here is to estimate the proportion of peer-reviewed academic-level literature which is available for free. Measuring how well Google Scholar fares at identifying a part of this is certainly an interesting exercise in itself, but it does not address our central question.

Finding that the tipping point has been reached in open access is certainly an important discovery. This means that the publishing industry is undergoing revolutionary change and at a pace much faster than anticipated, in large part because previous measures of OA availability proved to be misleading. This means that aggressive publishers such as Springer are likely to gain a lot in the redesigned landscape, whereas those attached to the old days are likely to suffer and to lose market share. The impression gained in carrying out this study and developing our OA 'trawler' is that the tool plaza is being moved to the beginning of the publishing process, away from the back-end of the process, and thus from the libraries and closer to researchers. Despite what several authors thought, and argued for, green OA only appears to move slowly, whereas Gold OA and hybrid toll before the process as opposed to toll after are in the fast lane. Efforts need to be made to characterise these changes.

If the toll plaza changes from the end of the process to the front-end, one category of workers is likely to be highly affected: the university and research centre librarian. Librarians have been highly affected already by the shift from paper to digital media and losing the responsibility of spending the large sum paid in journal subscriptions will certainly create another large dent in their traditional sphere of responsibilities. If the tool plaza is just moved, it means that researchers will have control over the toll.

The market power will shift tremendously from the tens of thousands of buyers that publishers' sales staff nurtured to the millions of researchers that will now make the atomistic decision of how best to spend their publication budget. Much has been said about the cost of publishing in gold and hybrid OA, but one has to place this in perspective. The cost of academic papers in the US is about $125,000 on average (HERD divided by number of papers by academia) so adding or including a $2,000 publication fee in this envelope is certainly going to break the bank. The question is rather whether the switch to a more atomistic market will reduce, augment or leave unchanged the negotiating power of publishers. One will have to stay tuned and watch the gales of creative destruction at play.

Table 4. Proportion of OA availability by country, 2008

Group	Country	Papers	Green & Hybrid		Gold		OA	
			Papers	%	Papers	%	Papers	%
EU27	Austria	107	32	30%	10	10%	42	40%
	Belgium	185	77	42%	5.0	3%	82	44%
	Bulgaria	24	8.9	37%	2.0	8%	11	45%
	Cyprus	5.1	1.1	21%	0.9	18%	2.0	39%
	Czech Republic	92	28	30%	5.7	6%	33	36%
	Denmark	145	48	33%	5.2	4%	54	37%
	Estonia	15	3.6	25%	2.9	20%	6.5	45%
	Finland	96	41	43%	5.5	6%	47	49%
	France	773	254	33%	41	5%	295	38%
	Germany	1,016	316	31%	59	6%	375	37%
	Greece	143	50	35%	11	8%	61	43%
	Hungary	74	21	28%	2.8	4%	24	32%
	Ireland	63	23	36%	3.7	6%	26	42%
	Italy	604	214	35%	32	5%	246	41%
	Latvia	6.9	4.5	65%		0%	4.5	65%
	Lithuania	21	9.2	44%	3.0	14%	12	58%
	Luxembourg	1.4	0.1	4%	1.0	72%	1.1	76%
	Malta	2.5	1.1	45%	0.4	15%	1.5	60%
	Netherlands	313	150	48%	14	4%	164	53%
	Poland	229	56	25%	27	12%	83	36%
	Portugal	82	31	37%	8.2	10%	39	47%
	Romania	64	25	39%	5.8	9%	31	48%
	Slovakia	43	14	32%	7.0	16%	21	49%
	Slovenia	34	10	29%	3.8	11%	14	40%
	Spain	549	162	30%	55	10%	217	40%
	Sweden	220	73	33%	11	5%	84	38%
	United Kingdom	1,147	465	41%	59	5%	523	46%
	Total EU27	**6,055**	**2,118**	**35%**	**383**	**6%**	**2,500**	**41%**
EFTA	Iceland	8	3	35%	1	13%	4	48%
	Liechtenstein	1	1	100%		0%	1	100%
	Norway	95	31	32%	9	10%	40	42%
	Switzerland	194	65	34%	14	7%	79	41%
	Total EFTA	**296**	**99**	**33%**	**25**	**8%**	**124**	**42%**
Candidate	Turkey	327	65	20%	50	15%	115	35%
	Croatia	38	16	41%	4	11%	20	52%
	Macedonia	3	1	42%	2	58%	3	100%
	Total Candidate	**368**	**82**	**22%**	**56**	**15%**	**138**	**37%**
	Israel	137	59	43%	4	3%	63	46%
Total ERA		**6,855**	**2,358**	**34%**	**467**	**7%**	**2,825**	**41%**
Others	United States	4,524	2,140	47%	220	5%	2,360	52%
	Japan	1,072	349	33%	76	7%	425	40%
	Canada	598	243	41%	29	5%	273	46%
	Brazil	450	89	20%	212	47%	301	67%

Source:Computed by Science-Metrix

Acknowledgments

This research has received support from the European Commission.

References

Abad-García, M. F., Melero, R., Abadal, E., & González-Teruel, A. (2010). Self-archiving of biomedical papers in open access repositories. *Autoarchivo de artículos biomédicos en repositorios de acceso abierto, 50*(7), 431-440. doi: 10.1371/journal.pbio.0040157.

Antelman, K. (2004). Do open-access articles have a greater citation impact? *College & Research Libraries*, 65(5), 372–382.

Arlitsch, K., & O'Brien, P. S. (2012). Invisible institutional repositories: Addressing the low indexing ratios of IRs in Google Scholar. *Library Hi Tech*, *30*(1), 60-81. doi: 10.1108/07378831211213210.

Bailey Jr, C. W. (2005). The role of reference librarians in institutional repositories. *Reference Services Review, 33*(3), 259-267.

Bilder, G. (2003). Ingenta's economic and technical models for providing institutional OA archives. *Information Services and Use, 23*(2-3), 111-112.

Björk, B. C., Welling, P., Laakso, M., Majlender, P., Hedlund, T., & Gudnason, G. (2010). Open Access To The Scientific Journal Literature: Situation 2009. PLoS ONE, 5(6). doi: 10.1371/journal.pone.0011273.

Carbone, P. (2007). Consortium negotiations with publishers - Past and Future. *LIBER Quarterly*, 17(2).

Chan, D. L. H., Kwok, C. S. Y., & Yip, S. K. F. (2005). Changing roles of reference librarians: The case of the HKUST Institutional Repository. *Reference Services Review, 33*(3), 268-282. doi: 10.1108/00907320510611302

Chan, L., Devakos, R., & Mircea, G. (2005). *Workshop 2: Implementing and filling institutional repositories introduction*, Leuven-Heverlee.

Charles, L., & Booth, H. A. (2011). An Overview of Open Access in the Fields of Business and Management. *Journal of Business and Finance Librarianship, 16*(2), 108-124. doi: 10.1080/08963568.2011.554786

Craig, I. D., Plume, A. M., McVeigh, M. E., Pringle, J., & Amin, M. (2007). Do open access articles have greater citation impact? A critical review of the literature. *Journal of Informetrics, 1*(3), 239-248. doi: 10.1016/j.joi.2007.04.001.

Dubini, P., Galimberti, P., & Micheli, M. R. (2010). Authors publication strategies in scholarly publishing. In *ELPUB 2010 International Conference on Electronic Publishing*, Helsinki (Iceland), 16-18 June 2010.

Gargouri, Y., Larivière, V. Gingras, Y. and Harnad, S. (2012). Green and Gold Open Access Percentages and Growth, by Discipline. In Archambault, É, Gingras, Y. and Larivière. V. (2012). *Proceedings of 17th International Conference on Science and Technology Indicators*, Montréal: Science-Metrix and OST.

Gentil-Beccot, A., Mele, S., & Brooks, T. C. (2010). Citing and reading behaviours in high-energy physics. *Scientometrics, 84*(2), 345-355. doi: 10.1007/s11192-009-0111-1

Harnad, S. (2003). The research-impact cycle. *Information Services and Use, 23*(2-3), 139-142.

Harnad, S. (2008). Waking OA's "slumbering giant": The university's mandate to mandate open access. *New Review of Information Networking, 14*(1), 51-68. doi: 10.1080/13614570903001322.

Harnad, S. (2012). Open access: A green light for archiving. *Nature, 487*(7407), 302. doi: 10.1038/487302b

Harnad, S., & Brody, T. (2004). Comparing the impact of open access (OA) vs. non-OA articles in the same journals. *D-Lib Magazine*, 10(6).

Henderson, I. (2013). Open-Access and Institutional Repositories in Fire Literature. *Fire Technology, 49*(1), 155-161. doi: 10.1007/s10694-010-0198-1

Houghton, J. W. (2010). Economic implications of alternative publishing models: Self-archiving and repositories. *LIBER Quarterly, 19*(3-4), 275-292.

Hu, C., Zhang, Y., & Chen, G. (2010). Exploring a New Model for Preprint Server: A Case Study of CSPO. *Journal of Academic Librarianship, 36*(3), 257-262. doi: 10.1016/j.acalib.2010.03.010

Kurek, K., Geurts, P. A. Th M., & Roosendaal, H. E. (2006). The split between availability and selection: Business models for scientific information, and the scientific process? *Information Services and Use, 26*(4), 271-282.

Lakshmi Poorna, R., Mymoon, M., & Hariharan, A. (2012). A study of select open access journals and their business models listed in DOAJ in the fields of civil and structural engineering. *Journal of Structural Engineering (India), 39*(4), 458-468. doi: 10.1371/ journal.pone.0020961.

Larivière, V., Archambault, É., Gingras, Y. & Vignola-Gagné, É. (2006). The Place of Serials in Referencing Practices: Comparing Natural Sciences and Engineering with Social Sciences and Humanities, *Journal of the American Society for Information Science and Technology*, 57(8), 997-1004. doi: 10.1002/asi.20349.

Miguel, S., Bongiovani, P. C., Gómez, N. D., & Bueno-de-la-Fuente, G. (2013). Prospect for Development of Open Access in Argentina. *Journal of Academic Librarianship, 39*(1), 1-2. doi: 10.1016/j.acalib.2012.10.002

Morris, S. (2003). Open Publishing: How publishers are reacting. *Information Services and Use, 23*(2-3), 99-101.

Nyambi, E., & Maynard, S. (2012). An investigation of institutional repositories in state universities in Zimbabwe. *Information Development, 28*(1), 55-67. doi: 10.1177/0266666911425264

Pelizzari, E. (2004). Academic authors and open archives: A survey in the social science field. *Libri, 54*(2), 113-122.

Prosser, D. (2003). Institutional repositories and Open access: The future of scholarly communication. *Information Services and Use, 23*(2-3), 167-170.

Regazzi, John. (2004). The Shifting Sands of Open Access Publishing, a Publisher's View. Serials Review, 30(4), 275-280. doi: http://dx.doi.org/10.1016/j.serrev.2004.09.010
Repanovici, A. (2012). Professional profile of digital repository manager. *Library Hi Tech News, 29*(10), 13-20. doi: 10.1108/07419051211294473
Rowland, F. et al. (2004). Delivery, management and access model for e-prints and open access journals. *Serials Review, 30*(4), 298-303. doi: 10.1016/j.serrev.2004.09.006
Sawant, S. (2012). Past and Present Scenario of Open Access Movement in India. *Journal of Academic Librarianship*. doi: 10.1016/j.acalib.2012.11.007
Swan, A. et al. (2005). Developing a model for e-prints and open access journal content in UK further and higher education. *Learned Publishing, 18*(1), 25-40. doi: 10.1087/0953151052801479
Swan, A., & Brown, S. (2004). Authors and open access publishing. *Learned Publishing, 17*(3), 219-224. doi: 10.1087/095315104323159649
Woutersen-Windhouwer, S. (2012). The Future of Open Access Publishing in the Netherlands: Constant Dripping Wears Away the Stone. *Journal of Academic Librarianship*. doi: 10.1016/j.acalib.2012.11.015