Global annual volume of peer reviewed scholarly articles and the share available via different Open Access options

Bo-Christer Björk; Annikki Roos; Mari Lauri

Bo-Christer Björk
Information Systems Science, Department of Management and Organization,
Swedish School of Economics and Business Administration
Arkadiankatu 22, 00100 Helsinki, Finland
bo-christer.bjork@hanken.fi;

Annikki Roos
National Public Health Institute
Mannerheimintie 166, Helsinki, Finland
annikki.roos@ktl.fi

Abstract

A key parameter in any discussions about the academic peer reviewed journal system is the number of articles annually published. Several diverging estimates of this parameter have been proposed in the past, and have also influenced calculations of the average production price per article, the total costs of the journal system and the prevalence of Open Access publishing. With journals and articles increasingly being present on the web and indexed in a number of databases it has now become possible to quite accurately estimate the number of articles. We used the databases of ISI and Ulrich's as our primary sources and estimate that the total number of articles published in 2006 by 23 750 journals was approximately 1 350 000.

Using this number as denominator it was also possible to estimate the number of articles which are openly available on the web in primary OA journals (gold OA). This share turned out to be 4.6 % for the year 2006. In addition at least a further 3.5 % was available after an embargo period of usually one year, bringing the total share of gold OA to 8.1%

Using a random sample of articles, we also tried to estimate the proportion of the articles published which are available as copies deposited in e-print repositories or homepages (green OA). Based on the article title a web search engine was used to search for a freely downloadable full-text version. For 11.3 % a usable copy was found. Combining these two figures we estimate that 19.4 % of the total yearly output can be accessed freely.

Keywords: scholarly publishing, articles, open access, article output

1. Introduction

"Open Access" means access to the full text of a scientific publication on the internet, with no other limitations than possibly a requirement to register, for statistical or other purposes. This implicitly means that Open access (OA) material is easily indexed by general purpose search engines. There are several widely quoted definitions on the net, for instance the Budapest Open Access Initiative (2002). For the scholarly journal literature in particular, OA can be achieved using two complimentary strategies: Gold OA means journals that are open access from the start, whereas green OA means that authors post copies of their manuscripts to OA sites on the web (Harnad et al 2004).

Since there are numerous different types of stakeholders involved in the scientific publishing value chain (Björk 2007), such as publishers, libraries and authors, with sometimes conflicting interests, a lot of what is being written about OA is strongly

biased either towards promoting open access or describing the dangers of open access to the scholarly publishing system. There has also been a discussion among OA advocates which of the two strategies (gold or green) is better. There is thus an urgent need for reliable figures concerning the yearly volumes of journal publishing, and the share of the yearly volume which is available as open access via different channels.

In most of the earlier discussions about the economy of journal publishing the focus has been on the number of journals, and costs such as the subscription cost have been mainly related to the individual title (i.e. European Commission 2006). This was natural due to the fact of the easy availability of subscription information for individual titles and for the handling of paper copies in libraries all over the world.

We argue that since the advent of the digital delivery for the contents and the electronic licensing of vast holdings of journal content ("the big deal") the focus should be more on the individual articles as the basic molecule of the journal system and any average costs should be related to the article. We also think that the ratio of open access articles to the overall number of articles published is a much more important indicator of the growing importance of OA than the number of OA titles compared to the number of titles in general.

2. Total number of articles published

A central hypothesis in this calculation was that the journals indexed by Thomson Scientific's (ISI) three citation databases (SCI, SSCI an AHCI) on average tend to publish far more articles per volume than the often more recently established journals not covered by the ISI, and that this should explicitly be taken into account in the estimation method.

We proceeded as follows. To estimate the total number of scholarly peer reviewed titles we used Ulrich's periodicals directory and conducted a search with the following parameters; *Academic/Scholarly*, *Refereed* and *Active*. In the winter 2007 this yielded a total of 23 750 journals.

For the case of the journals indexed by the ISI it was possible to extract the total number of articles published in the last completed year (2006) by conducting a search in the Web of Science (WoS). A general search was done covering all three indexes (Science Citation Index Expanded, Social Sciences Citation Index and the Arts and Humanities Citation Index). The parameters were set as follows; Publication Year = 2006, Language = All languages, Document type = Article. Since the system has a limitation in the number of items shown of 100 000 it was not possible to directly get the total number of indexed articles. The problem was solved by systematically going through the alphabet by setting the Source Title as A*, B*, C* etc. This worked well for all other letters, for which the total number was less than 100 000, except for A an J. For the letter A more detailed search on AA*, AB* etc was enough, for J we had to go down to the level of Journal of A*, Journal of B* etc. The total number of articles we arrived at in this way was 966 384.

ISI as a rule only indexes peer reviewed journals, but with at least one notable exception, the "lecture notes in…" series published by Springer, which publishes conference proceedings in computer science and mathematics in book form. By doing

a search using the above as Source Title we got the number of articles published in this series which was 20 484. Subtracting this from the total results in the final number of ISI articles of **945 900**.

If we know the exact number of titles that the ISI tracked in the WoS in 2006, we can easily derive the average number of articles published annually per title. Since we didn't have access to exact figures from ISI we had to go a roundabout way to estimate this figure. One indication is given by the number of journals included in the Journal Citation Reports. When searched from Ulrich's and defining *Journal Citation Reports* (JCR) as a further search criterion, the result is 6 877 titles. For one reason or another, the search directly from JCR for 2006 gives more journals: 6 166 titles indexed in SCI and 1 768 in SSCI. AHCI journals are not included in the Journal Citation Reports. We can, however, estimate the number of titles by assuming that AHCI journals on average publish as many articles per year as SSCI journals (53.1) which would result in additionally 532 titles. Summing these up, we would get 8 466 titles. Using these numbers as a base, we are able to estimate the average number of articles published in journals indexed in WoS by ISI as **111.7** per title. This can for instance be compared to the figure of 123 articles per year for 6 771 US publishers reported by Tenopir and King (2000).

The number of titles indexed in the WoS is probably slightly higher than our estimate for a couple of reasons. The main reason is a time lag between the inclusion in the indexes and the first journal citation report produced for a specific journal. According to ISI (Horky 2008) the number of titles indexed in the citation databases at the end of the year 2007 was 9 190 journals. In the beginning of 2008 according to ISI's web-pages the number of journals had risen to 9 300. This would indicate, assuming that the number of journals indexed rises steadily every year, that the number would have been somewhere between our estimate and this information. However, we have chosen to use our earlier mentioned estimate (8 466) because the number of titles does not influence the number of ISI-articles which we have obtained separately. It does affect our estimate of the number of non-ISI journals since these are obtained by subtraction (see text below). Since we have estimated these to have a much lower number of articles published per year the effects of a possible mistake in our number of ISI-titles of 1 % would be only around 0.2 % in the total number of articles.

Taking as a starting point the number of total titles 23 750 and the number of titles indexed by the ISI 8 466 we arrive by subtraction at a number of titles not indexed by the ISI of 15 284. In order to arrive at a total number of articles we now need to estimate how many articles these journals publish on average per year. This was done using a statistical sample of journals. The basis was Ulrich's database from which a statistical sample of 250 journals was taken. We set the search so that we only chose journals that have an on-line presence. This might statistically result in a slight bias, but was the only practical way we could study the publication volumes of the journals in the sample. We then extracted the number of articles published in 2006 until we had data for 104 journals (Journals in the original sample which were indexed by the ISI or for which the number of articles could not be found were discarded). In this group the average number of articles published was **26.2**, which as we had suspected was considerably lower than for ISI indexed journals. Five of the journals had published no articles and the journal with the highest output had published 225 articles. Multiplying 26.2 by 15 284 results in an estimate of articles published in

2006 of 400 440. Adding the figures for ISI brings the estimate of the total number of peer reviewed articles to **1 346 000** (rounded off) with 70 % covered by the ISI.

In their answer to a UK House of Commons committee Elsevier in 2004 estimated that some 2000 publishers in STM (Science, Technology and Medicine) publish 1.2 million peer reviewed articles annually (Elsevier 2004). Taking into account publishing in the social sciences and the humanities our estimate seems to be well in line with these figures.

3. Share of OA publishing

In policy discussions concerning Open Access publishing a very important question is "what share of all scientific articles is available openly". For a given year, in our case 2006, this concerns both articles directly published as open access (the so-called gold route in OA jargon), and articles published in subscription based journals, but where the author has deposited a copy in a subject-based or institutional repository (green route).

It is easier to estimate the number of gold route articles. For the case of copies in repositories, the evidence is much more scattered, and there is the additional difficulty of checking the nature of the copies (copy of manuscript submitted, personal copy of approved manuscript or replica of published article).

3.1. Gold

To estimate the number of articles directly available as OA in 2006 the Directory of Open Access journals (DOAJ) would at first sight seem to be the natural entry point. At the time of checking the directory listed 2 961 journals. Using the directory it is easy to go directly to the web pages of a journal and manually count the number of articles published. One problem is, however, that DOAJ states as inclusion criteria that journals are quality controlled by peer review or editorial quality control. When we searched Ulrich's for our earlier analysis, we only included journals which had self-reported as refereed (23 750 titles). If we relaxed that criterion and only required a journal to be active and scholarly/academic a search in Ulrich's yields 60 911 titles. The corresponding figures if the additional criterion of open access was defined were 1 735 refereed and 2 690 scholarly/academic in all. The latter figure is, as could be expected, quite close to the DOAJ total. For these reasons we decided to use Ulrich's as an entry point, concentrating on the 1 735 journals listed as refereed and open access. In doing the actual counting we tried as far as we could, based on the tables of contents on the web, to only include research articles, excluding editorials etc. This is in line with our earlier use of ISI where we concentrated on the article category only.

There are a handful of major OA publishers, Public Library of Science (PLoS), BioMed Central, Hindawi and Internet Scientific Publications (ISP) which use article charges or other means to fund their operations. We counted their articles separately, since they have some high-volume journals. All 7 PLoS journals are listed in Ulrich's as peer reviewed. Of the 176 BioMed Central journals listed in DOAJ 172 are also listed in Ulrich's as scholarly and 139 as refereed.

For OA journals by other publishers, often published on university web sites using an open source mode of operation with neither publication charges nor subscriptions, we again used a sampling technique. The starting point for this was the figure from Ulrich's of 1 735 OA titles in total from which we subtracted the number of titles operated by the four publishers listed above resulting in 1 487 titles. A selection of 100 journals was made from this set and the number of research articles was counted from the tables of contents on their web sites. This resulted in an estimate of the mean number of articles published per year of 34.6. Table 1 shows our calculation of the number of OA titles and the number of articles published in 2006. We estimated the total number to 61 313 and this represented 4.6 % of all articles published in 2006.

Table 1. Number of OA titles and articles in 2006

	Peer reviewed titles (Ulrich's)	Articles 2006
PLOS	7	881
Biomed Central	139	6 589
Hindawi	44	1 643
ISP	58	737
Other OA journals	1 487	51 465
SUM	1 735	61 313

Our figures can be compared to a number of earlier studies. Regazzi (2004) used a similar sampling method to study the journals listed in DOAJ in 2003 and 2004 and found a drop in the estimated total number of articles from 25 380 to 24 516, indicating an overall share of 2 % STM articles. He notes that OA journals on average publish far fewer articles (30 on average) than established journals, and quotes an average of 103 for ISI tracked STM journals and 160 for the 1 800 titles of Elsevier. We have also ourselves earlier studied this number through a web survey to the editors of open access journals and then obtained a rather lower figure of 16 articles per year (Hedlund, Gustafsson and Björk 2004).

In a white paper on open access publishing from Thomson Corporation (Mc Veigh 2004), the owner of ISI, numbers are given for open access articles included in the Science Citation Index. The text indicates that first the OA publishers were determined from the ROMEO database on publisher OA policies after which the articles were counted. The number of OA articles in SCI in 2003 was 22 095 out of a total of 747 060. Thus roughly 3.0 % of all articles in ISI's Science Citation Index would have been open access in that year.

3.2. Delayed and hybrid OA

In addition to pure gold OA publishing there are two additional routes which could be worth studying. These are the open publishing of individual articles in otherwise closed journals using a separate fee (sometimes labelled open choice) and delayed open access publishing of whole journals. The important thing is that in both these options the version accessed is the original publication, at the publishers website, the only difference is that the access restrictions have been lifted for either a single article, or for articles that have been published before a specific date.

All of the biggest publishers, Springer, Taylor & Francis, Blackwell, Wiley and Elsevier provide the option of freeing individual articles against a fee for a wide spectrum of journals (see Morris 2007). It is typical that this opportunity is offered to a sample of the journals in a publisher's collection. Oxford University Press is an example of a publisher which has been among the first hybrid providers and Karger is an example of a publisher which offers "Author's Choice" to all of its journals. There are no systematic studies on how commonly the open choice option has been chosen by authors but so far it appears to be rather low. We chose not to do any calculations of our own, since this would be very labor-intensive due to the scattering of relatively few articles among a vast number of titles.

Delayed open access is more common among society publishers than commercial publishers. A good example of an individual journal practicing delayed OA is Learned Publishing, the articles in which become OA roughly one year after publishing. A lower bound for an estimate of the prevalence of delayed OA can be obtained via the web portal of HighWire Press, which hosts the e-versions of currently 1 080 journals from over 130 mostly non-commercial publishers. Only a small number of the journals (43) are fully open access from the start but of the total of 4.6 million articles 1.8 million are freely available. The fully open access ones are such that the print version is subscription-based but the online version is free.

A search in the database for articles posted during 2006 results in 219 224 hits. This figure may not exactly coincide with the number or articles formally published during that same year and some caution is in order regarding the fact that some of the serials in HighWire Press should not be classified as fully refereed scholarly journals. Of the 1080 HighWire journals 277 (as of January 2008) offer direct or delayed open access. Table 2 lists the numbers in different delay categories as well as an estimate of the total number of articles. The latter has been made assuming that the average number of articles for these is the same as for all the journals in the HighWire portal.

Table 2. OA articles published electronically by HighWire Press.

Delay	No of journals	% of all HW journals	Estimated number of articles
Direct OA	43	4,0	8 700
1-6 months	27	2,5	5 481
7-12 months	190	17,6	38 567
Over 12 months	17	1,6	3 451
Delayed in total	234	21,7	47 499

Thus comparing this to the total number of articles published in 2006 the share of delayed OA can be estimated to at least 3.5 %, bringing the sum of direct and delayed gold OA to 8.1 %.

From the viewpoint of readers hybrid ("open choice") and delayed open access are less useful than full and instant open access on the title level in "current awareness" reading, where academics track what is being published in a few essential journals either by getting a paper copy or an e-mail table-of-content message. This type of

information activity is called “monitoring” in Ellis’s model of information-seeking behaviour (Ellis 2005). Hybrid and delayed open access help more in cases where a reader tries to access a given article based on a citation (called “chaining” in Ellis’s model).

3.3. Parallel publishing of copies (green)

It is much more difficult to estimate the prevalence of green OA than gold OA. Copies of articles published in referee journals are scattered in hundreds of different repositories as well as in even more numerous homepages of authors. There is also the issue of the actual existence of a digital copy on some server versus how easy it is to find it using the most widely used web search engines.

For the purposes of this article, we take the pragmatic view that unless you get a hit in Google (or Google Scholar) using the full title of an article, a copy “does not exist”. This is both because a copy which cannot be found this way is very difficult to find for a potential reader and because the best systematic way of measuring the proportion of “green articles” is via systematic search on article titles using a popular search engine, such as Google.

An additional complication is that the full text copy found may differ quite substantially from the final published version. It can in the best of cases be an exact copy of the published file (usually PDF) but it can also be a manuscript version from any stage of the submission process. The most useful version is often labelled “accepted for publication” and sometimes includes also changes resulting from the final copy-editing done by the publisher’s technical staff, sometimes not. The layout and page numbering is also usually different from the final published version. Most publishers who allow posting of a copy of an article in an e-print repository allow posting of this so-called “personal version”. In addition, some researchers also upload earlier manuscript versions, often called preprints, but this is not as common except for certain disciplines such as physics.

In order to estimate the green route to open access we selected a random sample of all peer reviewed articles published in 2006. The entry point was again Ulrich’s, out of which we took a sample of both journals listed in ISI Web of Science as well as those not listed there. The sample was proportional so that the number of articles from ISI corresponded roughly to the share of ISI in the total number of articles (it included 200 articles in ISI journals and 100 articles in non-ISI journals). A spreadsheet listing the title of the article, the three first authors and the name of the journal was created from the sample. A search was then conducted in Google systematically using the name of the article and in the second hand the writers’ names, using a computer which had Internet access but no access to our university intranet which would automatically allow access to the journals we subscribe to. (We first tried also Google Scholar but we dropped it after a while since we noticed that the search results turned out to almost identical). In order to keep the workload manageable and follow the viewpoint of an average searcher, who does not want to spend too much time and energy, we only searched the 10 first hits, which also is what you usually see on the first screen. If we got a hit which was not on the journal’s own website and which included a full text file containing a document available without subscription, that seemed to fulfil the criteria, a copy was downloaded and saved.

The last check was performed by comparing the obtained copy to the published official version which we obtained separately via our own university website or the website of the publisher. This was in order to see that the copy was close enough to the original article. Out of the 35 copies we studied we had subscribed access to 32 and were able to do the comparison, for the remaining three we assumed the copies to be usable. Two of the copies studied turned out to differ significantly in content from the original, and were therefore discarded.

The results concerning copies in repositories were very similar for ISI-indexed journals (11%) and the other journals (12 %) bringing the weighted average to 11.3 %. The spread between different formats and different types of repositories is shown in the table below, but the absolute numbers are so small per category that it is difficult to generalize to the whole target population. Table 3 shows the percentage of green OA-versions and their popularity:

Table 3.The frequency of OA copies of different kinds.

Type of site	Type of copy			
	Exact copy	Personal version	Other version	All
Subject based repository	0.7	2.3	0.3	3.3
Institutional repository	4.7	3.0	0.0	5.0
Author's home pages	1.7	1.3	0.0	3.0
All	7.0	4.0	0.3	11.3

We found no case of overlaps of the same article being both published as gold OA on a publisher's website and with a copy in a repository. Thus the figures for green OA can be added to our earlier estimates for gold OA (8.1 %) to get the total OA availability of 19.4%.

We were of course also able to check the direct gold availability of the articles in the sample. For the articles in ISI journals the percentage was 15 but for non-ISI articles an astonishing 35 %, compared with our earlier figures of 8.1 %. The reasons for this can be twofold. Firstly we were in practice restricted in producing the sample to journals which at least have tables of content freely available on the net. Our experience in producing the sample, in terms of how many candidate journals we had to disqualify because of a lacking web presence, indicated that for ISI-listed journals the availability of web tables of contents is nowadays rather high, whereas for non-ISI journals the percentage is much lower. Unfortunately we did not keep exact records when we produced our sample, which could have helped in correcting the estimate taking this factor into account. Secondly there might be a random element in this calculation, which of course could be reduced by increasing the sample size. All in all we believe our earlier estimate of gold availability to be more reliable.

4. Conclusions and discussion

We have estimated in this study that the amount of scientific articles published in 2006 was 1 346 000. Our hypotheses about the difference in the number of articles

published per title in the titles indexed by ISI and non-ISI-titles appeared to be correct. The non-ISI journals published on average 26.7 articles per title and the ISI-journals 111.7 articles. 4.6 % from the yearly article output appears in the Golden OA journals and at least 3.5 % is open after a delay period. 11.3 % of articles are openly available in repositories and for example on personal web pages. Altogether the amount of openly available articles from the yearly output is 19.4 %.

The different elements in our calculation differ in terms of accuracy. The total number of articles included in the indices of the ISI should be very accurate, provided that we have searched the database in a correct way. Also the total number of journals tracked by the ISI in a given year is reasonably accurate. The total number of peer reviewed scholarly journals is much more difficult to estimate accurately. Ulrich's database is the best tool available for this purpose, but its coverage is not 100 %, and there are some inactive journals included in the category. On the other hand if we organise the total journal market according to the number of yearly articles per title we get a distribution with a few very high volume titles and many journals with few articles. It is very likely that journals which are not listed in Ulrich's publish rather few articles per annum, and thus their contribution to the total volume of article is rather marginal.

It is also more likely that Ulrich's coverage of journals published in the Anglo-Saxon countries is more comprehensive than journals published in non-English speaking countries and in particular in languages other than English. It is also impossible to draw a clear border line between journals practicing full peer review and journals where the editors check the content of the submission. In this respect we just have to trust the self-reporting of journals to Ulrich's data base. Also we have excluded conference proceedings produced using a referee procedure, since it would be very difficult to find data about these. The one notable exception is the Springer Lecture Notes series, but we chose to exclude it from our calculations.

An interesting study of the growth of Open Access and the effect of open vs. closed access on the number of citations has been carried out by Hajjem, Harnad and Gingras (2005). They used a web robot to search for full texts corresponding to the citation metadata of 1.3 million articles indexed by the ISI from a 12 year time period (1992-2003), in particular focusing on differences between disciplines in the degree of open availability and in the citation advantage provided by OA. Articles published in OA journals were excluded and their results thus concern articles published in subscription-based journals where the author (or a third party) has deposited a copy on any web site which allows full text retrieval for web robots. According to the study the degree of green OA varied from 5-16 % depending on the discipline, but from our viewpoint the most important figure was that for the total of 1.3 million articles OA full text copies could be found for 12 %. This included both direct replicas, the author's accepted manuscripts after the review ("personal version") and submitted manuscripts ("preprint"), since it can be assumed that the robot could not distinguish between these if the title and author have remained unchanged.

All in all we believe our estimates to be more accurate than the estimates that have been presented earlier in different contexts. We have defined our method in detail and the estimate can easily be replicated and/or adjusted by other researchers in later years.

Acknowledgements:

This study was partly financed by the Academy of Finland, through the research grant for the OACS project (application no. 205993). We would also like to thank Piero Ciarcelluto for his assistance in the data gathering phase.

REFERENCES:

Björk B-C. 2007. A model of scientific communication as a global distributed information system**.** *Information Research,* Vol. 12(2) paper 307. Available at http://InformationR.net/ir/12-2/paper307.html

Budapest Open Access Initiative. 2002. *http://www.soros.org/openaccess/read.shtml*

Elsevier. 2004. Responses to the questions posed by the Science and Technology Committee, *Document submitted to the UK House of Commons Select Committee on Science and Technology by Elsevier on 12 February 2004.* Available at: http://www.elsevier.com/authored_news/corporate/images/UK_STC_FINAL_SUBMISSION.pdf

Ellis, D. (2005) Ellis's model of information-seeking behaviour. In Fisher K.E. et al. (eds.) *Theories of information behaviour*. Medford : Information Today. pp. 138-142,

European Commission. 2006. *Study on the economic and technical evolution of the scientific publication markets in Europe.* Brussels: European Commission. Directorate General for Research. Available at: http://ec.europa.eu/research/science-society/pdf/scientific-publication-study_en.pdf

Hajjem C., Harnad S. and Gingras Y. 2005. Ten-Year Cross-Disciplinary Comparison of the Growth of Open Access and How it Increases Research Citation Impact. *IEEE Data Engineering Bulletin* Vol. 28(4) pp. 39-47. Available at: http://eprints.ecs.soton.ac.uk/11688/

Harnad S., Brody T., Vallières F., Carr L., Hitchcock S., Gingras Y., Oppenheim C., Stamerjohanns H. and Hilf ER. 2004. The Access/Impact Problem and the Green and Gold Roads to Open Access. *Serials Review,* Vol. 30(4), pp. 310-314.

Hedlund T., Gustafson T. and Björk B-C. 2004. The Open Access Scientific Journal: An Empirical Study. *Learned Publishing,* Vol. 17(3), pp. 199-209.

Horky, David 2008. E-mail from David Horky, Thomson Scientific, in the 17th of Jan 2008.

Mc Veigh ME. 2004. *Open Access Journals in the ISI Citation Databases: Analysis of Impact Factors and Citation Patterns* – A Citation study from Thomson Scientific. http://scientific.thomson.com/media/presentrep/essayspdf/openaccesscitations2.pdf

Morris S. 2007. Mapping the journal publishing landscape: how much do we know? *Learned Publishing,* Vol.20(4), pp. 299-310.

Regazzi J. 2004. The Shifting Sands of Open Access Publishing, a Publisher's View. *Serials Review* Vol. 30, pp. 275-280.

Tenopir, C. & King, D. (2000). *Towards electronic Journals – realities for scientists, librarians and publishers*, Washington D. C.: Special Libraries Association.