informationresearch

VOL. 14 NO. 1, MARCH, 2009

Contents | Author index | Subject index | Search | Home

Scientific journal publishing: yearly volume and open access availability

Bo-Christer Björk, Annikki Roos and Mari Lauri
Information Systems Science, Department of Management and Organisation,
Hanken School of Economics, Arkadiankatu 22, 00100 Helsinki, Finland

Abstract

Introduction. We estimate the total yearly volume of peer-reviewed scientific journal articles published world-wide as well as the share of these articles available openly on the Web either directly or as copies in e-print repositories.
Method. We rely on data from two commercial databases (ISI and Ulrich's Periodicals Directory) supplemented by sampling and Google searches.
Analysis. A central issue is the finding that ISI-indexed journals publish far more articles per year (111) than non ISI-indexed journals (26), which means that the total figure we obtain is much lower than many earlier estimates. Our method of analysing the number of repository copies (green open access) differs from several earlier studies which have studied the number of copies in identified repositories, since we start from a random sample of articles and then test if copies can be found by a Web search engine.
Results. We estimate that in 2006 the total number of articles published was approximately 1,350,000. Of this number 4.6% became immediately openly available and an additional 3.5% after an embargo period of, typically, one year. Furthermore, usable copies of 11.3% could be found in subject-specific or institutional repositories or on the home pages of the authors.
Conclusions. We believe our results are the most reliable so far published and, therefore, should be useful in the on-going debate about Open Access among both academics and science policy makers. The method is replicable and also lends itself to longitudinal studies in the future.

CHANGE FONT

Introduction

It is important to begin this paper with two definitions that are central to the entire discourse. By *Scientific Journal Paper* we mean a paper describing scientific research results, which has undergone some form of anonymous peer-review and which is published in a regularly appearing serial, usually by a third party publisher and not by the university of the author. Journals fall into the science, technology and medicine category as well as social science and the humanities. An alternative term often used is scholarly journals, but we have chosen the term scientific to cover all these subjects. For instance Tenopir and King (2000) sometimes speak of scientific scholarly journals in their influential book.

Papers are typically 3,000 to 10,000 words in length and are written following long-established conventions concerning style, referencing, tables of content etc. Other types of scientific publication include conference papers, book chapters, books and reports. Journal publishing is the most common form of dissemination of new research results, in particular in science and medicine. In some scientific domains, such as computer science, conference publishing is quite important and, in the humanities, book publishing is an important channel. Our analysis deals only with peer-reviewed papers published in journals.

Compared with the other types of scientific publication, journal papers are comparatively easier to obtain, even years after publication, because of the large holdings of journals by university libraries. Today, the vast majority of recent journal papers is also available electronically. Most of the larger universities have licenses offering access to all the titles of major publishers (e.g. Science Direct) and many publishers also offer pay-on-demand services for the purchase of individual papers.

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

As there are numerous different types of parties involved in the scientific publishing value chain (Björk 2007), such as publishers, libraries and authors, with sometimes conflicting interests, much of what is written about open access is strongly biased either towards promoting open access or describing the dangers of open access to the scientific publishing system. There has also been a discussion among open access 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 (e.g. European Commission 2006). This was natural because 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 papers as the basic molecule of the journal system and that any average costs should be related to the paper. We also think that the ratio of open access papers to the overall number of papers published is a much more important indicator of the growing importance of open access than the number of open access titles compared to the number of titles in general.

Total number of papers published

A central hypothesis in this calculation was that the journals indexed by Thomson Scientific's (ISI) three citation databases (Science Citation Index, Social Science Citation Index and Arts and Humanities Citation Index) on average tend to publish far more papers in a 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 scientific peer-reviewed titles we used *Ulrich's Periodicals Directory* and conducted a search with the following parameters; *Academic/Scholarly*, *Refereed* and *Active*. In 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 papers 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 papers. 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 and 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 papers we arrived at in this way was 966,384.

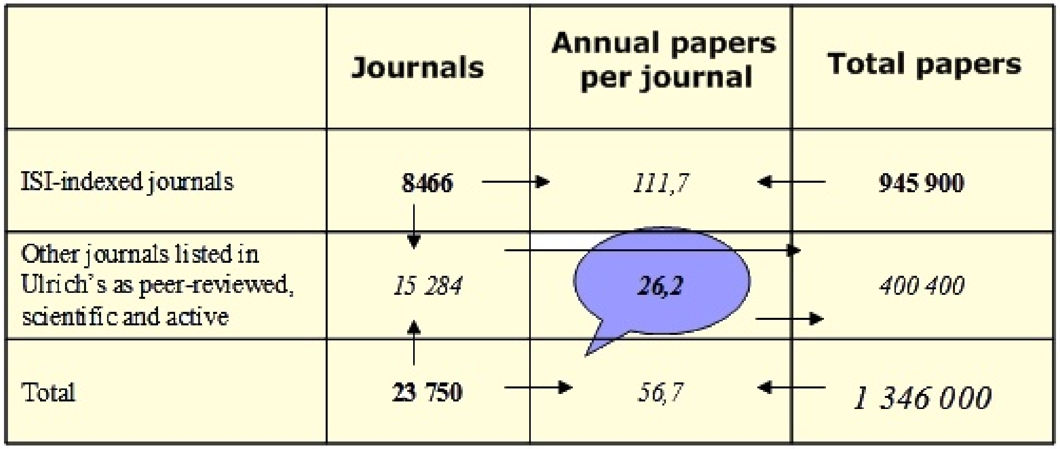
ISI, as a rule, only indexes peer-reviewed journals, but with at least one notable exception, the *Lecture Notes in...* series published by Springer, which publishes conference proceedings in computer science and mathematics in book form. By doing a search using the above as Source Title we got the number of papers published in this series, which was 20,484. Subtracting this number from the earlier total leads to a final number of 945,900 ISI papers.

If we know the exact number of titles that the ISI tracked in the Web of Science in 2006, we can easily derive the average number of papers published in a year by each title. As we did not have access to exact figures from ISI we had to find 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* as a further search criterion, the result is 6,877 titles. For one reason or another, the search directly from Journal Citation Reports for 2006 gives more journals: 6,166 titles indexed in Science Citation Index and 1,768 in Social Science Citation Index. Arts and Humanities Citation Index journals are not included in the Journal Citation Reports. We can, however, estimate the number of titles by assuming that Arts and Humanities Citation Index journals on average publish as many papers a year as Social Science Citation Index journals (53.1) which would result in an additional 532 titles. Adding these together, we would get 8,466 titles. Using these numbers as a base, we are able to estimate the average number of papers published in journals indexed in Web of Science by ISI as 111.7 per title. This can, for instance, be compared to the figure of 123 papers per year for 6,771 US publishers reported by Tenopir and King (2000).

The number of titles indexed in the Web of Science is probably slightly higher for two 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 (Personal communication from David Horky, Thomson Scientific, 17th of January, 2008) the number of titles indexed in the citation databases at the end of the year 2007 was 9,190 journals. At the beginning of 2008, according to ISI's Web-pages, the number of journals had risen to 9,300. Assuming that the number of journals indexed rises steadily every year, this would indicate 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-papers 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). As we have estimated these to have a much lower number of papers published in a 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 papers.

Taking as a starting point the total number of titles as 23,750 and the number of titles indexed by the ISI as 8,466 we arrive by subtraction at the number of titles not indexed by the ISI as 15,284. In order to arrive at a total number of papers we now need to estimate how many papers these journals publish yearly, on average. This was done using a statistical sample of journals. The basis was *Ulrich's* database from which a sample of 250 journals was taken. We set the search so that we chose only journals that have an online 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. Then we extracted the number of papers published in 2006 until we had data for 104 journals (journals in the original sample that were indexed by the ISI or for which the number of papers could not be found were discarded). In this group the average number of papers published was 26.2, which, as we had suspected, was considerably lower than for ISI-indexed journals. Five of the journals had published no papers and the journal with the highest output had published 225 papers. Multiplying 26.2 by 15,284 results in an estimate of papers published in 2006 of 400,440. Adding the figures for ISI brings the estimate of the total number of peer-reviewed papers to 1,346,000 (rounded off) with 70% covered by the ISI.

In their answer to a UK House of Commons committee in 2004, Elsevier estimated that some 2,000 publishers in science, technology and medicine publish 1.2 million peer-reviewed papers annually (U.K. *Parliament*... 2004). Taking into account publishing in the social sciences and the humanities our estimate seems to be well in line with these figures.



	Journals	Annual papers per journal	Total papers
ISI-indexed journals	**8466**	*111,7*	**945 900**
Other journals listed in Ulrich's as peer-reviewed, scientific and active	*15 284*	***26,2***	*400 400*
Total	**23 750**	*56,7*	*1 346 000*

Table 1: Estimated total number of journal papers published in 2006.
The figures in bold have been extracted from the two databases used (ISI and *Ulrich's*). The highlighted figure in the centre was determined based on counting papers for a statistical sample of non-ISI journals with tables of content on the Web. The figures in italics result automatically from the other parameters by simple arithmetic operations.

Share of open access publishing

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

It is easier to estimate the number of *gold* route papers. In 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).

Gold

To estimate the number of papers directly available as open access in 2006, the *Directory of Open Access Journals* 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 count the number of papers published. One problem, however, is that the *Directory of Open Access Journals* 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 and/or 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 and/or academic in total. The latter figure is, as could be expected, quite close to the *Directory of Open Access Journals* 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 and based on the tables of contents on the Web, to include only research papers and to exclude editorials, book reviews, etc. This is in line with our earlier use of ISI where we concentrated on the article category only.

There is a handful of major open access publishers, Public Library of Science, BioMed Central, Hindawi and Internet Scientific Publications, which use author charges or other means to fund their operations. We counted their papers separately since they have some high-volume journals. All seven Public Library of Science journals are listed in *Ulrich's* as peer-reviewed. Of the 176 BioMed Central journals listed in the *Directory of Open Access Journals*, 172 are also listed in *Ulrich's* as scholarly and 139 as refereed.

For open access 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 open access titles 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 papers was counted from the tables of contents on their Web sites. This resulted in an estimated mean of 34.6 papers published annually. Table 2 shows our calculation of the number of open access titles and the number of papers published in 2006. We estimated the total number of open access papers to be 61,313 and this represented 4.6% of all papers published in 2006.

	Peer reviewed titles (*Ulrich's*)	**Papers 2006**
Public Library of Science	7	881
Biomed Central	139	6,589
Hindawi	44	1,643
Internet Scientific Publications	58	737
Other open access journals	1,487	51,465
Total	1,735	61,313

Table 2: Number of open access titles and papers in 2006

Our figures can be compared to a number of earlier studies. Regazzi (2004) used a similar sampling method to study the journals listed in the *Directory of Open Access Journals* in 2003 and 2004 and found a drop in the estimated total number of papers from 25,380 to 24,516, indicating an overall share of 2% science, technology and medicine papers. He notes that open access journals on average publish far fewer papers (30 on average) than established journals and quotes an average of 103 for ISI-tracked science, technology and medicine 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 papers a year (Hedlund *et al.* 2004).

In a white paper on open access publishing from Thomson Corporation the owner of ISI (McVeigh 2004), numbers are given for open access papers included in the Science Citation Index. The text indicates that first the open access publishers were determined from the ROMEO database (Sherpa/Romeo 2008) on publisher open access policies after which the papers were counted. The number of open access papers in Science Citation Index in 2003 was 22,095 out of a total of 747,060. Thus, roughly 3.0% of all papers in ISI's Science Citation Index would have been open access in that year.

Delayed and hybrid open access

In addition to pure *gold* open access publishing there are two additional routes worth studying. These are the open publishing of individual papers 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 publisher's Website, the only difference is that the access restrictions have been lifted for either a single article, or for papers 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 papers 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 the figures appear to be rather low. We chose not to do any calculations of our own, since this would be very labour-intensive because of the scattering of relatively few papers 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 open access is Learned Publishing, the papers in which become open access roughly one year after publishing. A lower bound for an estimate of the prevalence of delayed open access can be obtained through the Web portal of HighWire Press, which currently hosts the e-versions of 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 sum of 4.6 million papers 1.8 million are freely available. The fully open access ones are such that the print version is subscription-based but the online one free.

A search in the database for papers posted during 2006 results in 219,224 hits. This figure may not exactly coincide with the number or papers 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 scientific journals. Of 1,080 HighWire journals, 277 (as of January 2008) offer direct or delayed open access. Table 3 lists the numbers in different delay categories as well as an estimate of the total number of papers. The latter has been made assuming that the average number of papers for these is the same as for all the journals in the HighWire portal.

Delay	No. of journals	% of all HighWire journals	Estimated number of papers
Direct open access	43	4.0	8,700
2-6 months	27	2.5	5,481
12 months	190	17.6	38,567
24 months or longer	17	1.6	3,451
Delayed in total	234	21.7	47,499

Table 2: Open access papers published electronically by HighWire Press.

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

From the viewpoint of readers hybrid (i.e., 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).

Parallel publishing of copies (*green* open access)

It is much more difficult to estimate the prevalence of *green* open access than *gold* open access. Copies of papers published in refereed journals are scattered in hundreds of different repositories as well as in even more numerous home-pages 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 paper, 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 papers* is by systematic search on article titles using 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 changes resulting from the final copy-editing done by the publisher's technical staff and sometimes does 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 papers published in 2006. The entry point was again *Ulrich's*, out of which we took a proportional sample of journals listed in ISI Web of Science and those not listed there. The sample was proportional so that the number of papers from ISI corresponded roughly to the share of ISI in the total number of papers (it included 200 papers in ISI journals and 100 papers 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 also the authors' names, using a computer with Internet access but no access to our university intranet, which would automatically allow access to the journals to which we subscribe. To keep the workload manageable and to follow the viewpoint of an average searcher, who does not want to spend too much time and energy on a search, we only searched ten 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, which 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 through 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 thirty-five copies we studied, we had subscription access to thirty-two 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 therefore were 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 for each category that it is difficult to generalize to the whole target population. Table 3 shows the percentage of *green* open acccess versions and their popularity:

Type of site	Type of copy			
	Exact copy	Personal version	Other version	All
Subject based repository	0.7	23	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

Table 3.The frequency of open access copies of different kinds.

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

We were, of course, also able to check the direct *gold* availability of the papers in the sample. For the papers in ISI journals, the percentage was 15 but for non ISI papers an astonishing 35%, to be compared to our earlier figures of 8.1%. The reasons for this difference can be twofold. Firstly we were in practice restricted when producing the sample to journals that 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 it 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.

Method discussion

In our study we used total numbers of registered papers or journals from third party databases, whenever available. This was the case, for instance, for the total number of papers indexed by the ISI in 2006 or the total number of journals registered in *Ulrich's* database. Concerning this type of data the main uncertainty is related to how well these databases cover the intended population. Since ISI is only a subset of the total set of papers and since ISI has relatively strict quality criteria we don't believe that the figures from ISI are problematic. Also the article count is exact for this part.

The issue of coverage is more problematic for *Ulrich's* database, which also provides the residual term in the total number of journals and is used as a basis for an estimate of about one third of the total number of papers, using sampling. It is, for instance, more likely that *Ulrich's* coverage of journals published in English-speaking countries is more comprehensive than journals published in non-English speaking countries and in particular in languages other than English. In order to study this further we examined the listings of scientific journals published in Finland, for which we have reliable data.

Ulrich's listed a total of ninety-seven academic or scholarly, active, refereed journals published in Finland. The Federation of Finnish Learned Societies has a list of 132 current Finnish serial publications. ISI in its turn listed seventeen Finnish journals. A closer look at these lists allowed us for the purpose of this study to exclude several of these as they proved to be monographic series, inactive or otherwise not applicable (for example, copies in parallel languages or publisher no longer Finnish). After excluding the inapplicable titles, we were left with 54 titles in *Ulrich's*, 38 in the Federation list and 14 in ISI. Out of these only eight were listed in all three sites. Seventeen were listed on both *Ulrich's* and the Federation list. Of those 21 titles listed by the Federation and not by *Ulrich's* 19 are mainly published in other languages than English. Thus, based on this one case, there seem to be inaccuracies in *Ulrich's* as well, but it is the best tool currently available to us. The journals not in *Ulrich's* were rather small and publish relatively few papers per annum. Therefore, their contribution to the total volume of article is rather marginal and at least partially offset by the fact that there are journals listed in *Ulrich's* that do not fit our definition of a scientific journal.

The problem of coverage was also the reason we preferred to use *Ulrich's* a basis for the figure of open access journals rather than the *Directory of Open Access Journals*. This is because *Ulrich's* has the criterion that journals should be refereed, whereas the *Directory*, on closer scrutiny, includes quite a lot of non-refereed journals.

A third coverage problem is related to the fact that we only used HighWire Press data to estimate the prevalence of delayed open access. It is likely that the figures we got underestimate the total number, but getting data by going through the journal pages of individual publishers would have been extremely time-consuming.

The second major source of uncertainty is statistical and related to the sampling methods and sample sizes we used for estimating the number of yearly papers, as well as the frequency of *green* open access availability and type of open access copy. Table 4. recapitulates the major data.

Parameter to be estimated	Population	Sample size	Estimated parameter	Error margin
Papers a year in non-ISI journals	15,284	104	26.2	+/- 2.2
Papers a year in open access journals	1,487	100	34.6	+/- 3.2
Green open access availability of papers	1,340,000	300	11.3%	+/- 0.4%
Exact *green* open access copy	151,000	35	62%	+/- 10.0%

Table 4: Sizes of population, sample sizes we used, estimated parameters and the corresponding margins of error with a 95% level of confidence.

In our calculations we assumed that we analysed journals that represented a simple random sample from a large population. Using a 95% level of confidence, we obtain the confidence interval as the parameter estimate plus or minus 1.96 standard errors.

Conclusions and discussion

In this study, we have estimated that the number of scientific papers published in 2006 was 1,346,000. Our hypotheses about the difference in the number of papers 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 papers per title and the ISI-journals 111.7 papers. Four-point-six percent from the yearly article output appears in the *gold* open access journals and at least 3.5% is open after a delay period. Eleven-point-three percent of the papers are openly available in repositories and for example on personal Web pages. Altogether the proportion of openly available papers from the yearly output is 19.4%. The situation is illustrated in Figure 1.

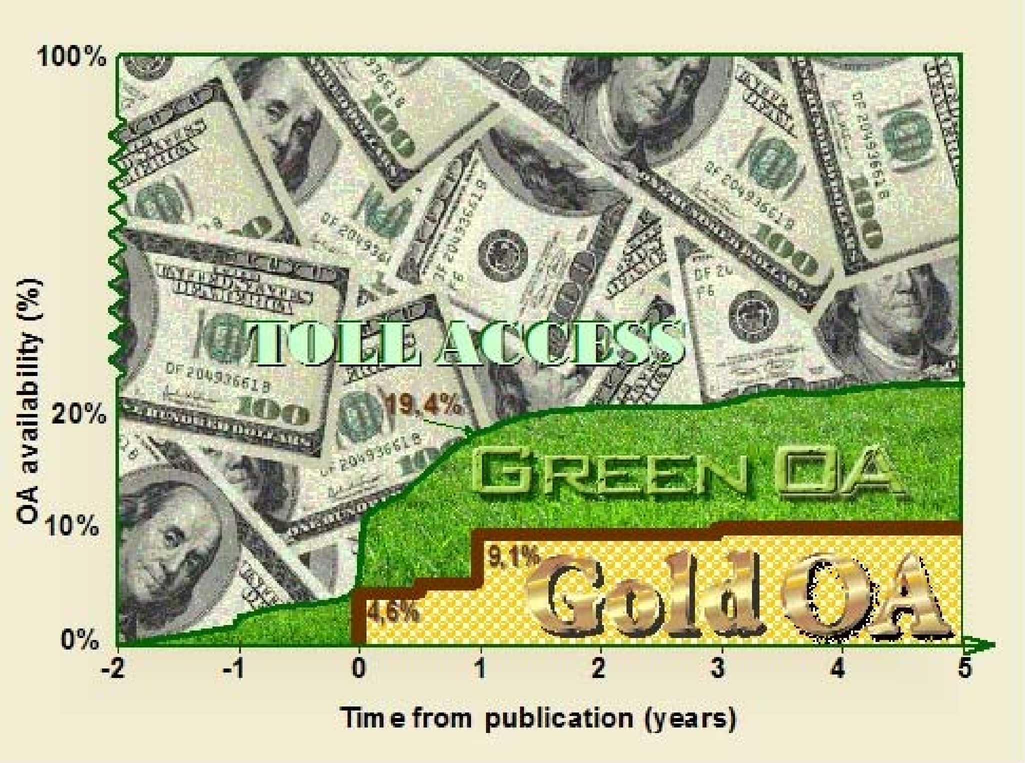


Figure 1: The availability of peer reviewed journal papers in open access mode as a function of time.

Some papers are available even before formal publishing as preprint or personal versions posted in e-print repositories.

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

The different elements in our calculation differ in terms of accuracy. The total number of papers included in the indices of the ISI should be very accurate, provided that we have searched the database in a correct way and the results can obviously be checked with the ISI. Also the total number of journals tracked by the ISI in a given year is information which can be verified.

The total number of peer reviewed scientific journals is much more difficult to estimate accurately. *Ulrich's* database is the best tool available for this purpose, but even *Ulrich's* is not very accurate, especially when it comes to small journals published in other languages than English. Furthermore, it appears that *Ulrich's* contains many monographic series and inactive journals. On the other hand if we organise the total journal market according to the number of papers a year per title we get a distribution with a few very high volume titles and many journals with few papers. It is very likely that journals not listed in *Ulrich's* publish fewer papers a year and, thus, their contribution to the total volume of papers is rather marginal. We still believe that it is our best starting point for the purpose of this study, as checking all national listings would be far too time-consuming and labor-intensive, even if all of them were accurate and readily available.

As an interesting point it can be noted that the comparison of the listings of Finnish academic journals also supports the general finding of this study that ISI-tracked journals publish on the average more papers than journals not included in the ISI. The Finnish academic journals listed in any of the three above-mentioned sites published on the average 20.0 papers a year whereas the 14 journals tracked by ISI had over double the number: an average of 47.2 papers.

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* database. Also we have excluded conference proceedings produced using a referee procedure, since it would be very difficult to find data about these.

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 *et al.* (2005). They used a Web robot to search for full texts corresponding to the citation metadata of 1.3 million papers indexed by the ISI from a 12 year time period (1992-2003), focusing, in particular, on differences between disciplines in the degree of open availability and in the citation advantage provided by open access. Papers published in open access journals were excluded and, consequently, their results concern papers published in subscription-based journals where the author (or a third party) has deposited a copy on any Web site that allows full text retrieval by Web robots. The degree of green open access 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 papers open access full-text copies could be found for 12%. This included both direct replicas, the author's accepted manuscripts after the review (or *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.

An area where the estimate could be useful is for any calculations of the average price, or the cost of publishing, a scientific peer-reviewed article. In the past many estimates have been made, usually based on cost data obtained from publishers. We propose a different method. In today's environment, an increasing proportion of journal income is through big consortia licence agreements (the so-called *big deals*). Publishers tend to treat information about their total subscription income as trade secrets. There is, however, a roundabout way to estimate the total global income to all publishers. In many countries, the ministries of education, associations or research libraries etc. produce statistics on the total expenditure on subscriptions and libraries. If reasonably accurate estimates can be produced for a few select countries, we can extrapolate this to the world market, by assuming that the share of these countries correlates with other factors such as share of authorship in ISI journals, research expenditure, GDP etc. Another factor to bear in mind is the share in journal income of subscriptions from the academic sector. Provided we get an estimate of the total revenue we can divide this by the number of papers published annually and thus get an estimate of the price per article

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 and Péter Majlender for their assistance and comments.

About the authors

Bo-Christer Björk is professor of Information Systems Science at the Hanken School of Economics in Helsinki, Finland. He is the editor-in-chief of the Journal of Information Technology in Construction (ITcon), an early Open Access journal published since 1996. He can be contacted at: Bo-Christer.Bjork@hanken.fi

Annikki Roos is the head of the Research Support Unit at the National Institute for Health and Welfare in Helsinki, Finland. She is also a doctoral student in Information systems science at the Hanken School of Economics, Helsinki. She can be contacted at: annikki.roos@thl.fi

Mari Lauri is an Amanuensis in the Institute of Development Studies at the University of Helsinki . She received her Bachelor's degree in Development Studies at the University of Gothenburg in Sweden and her Master's in International Business from the Helsinki School of Economics. At the time of writing this article she was working for the Hanken School of Economics. She can be contacted at: mari.lauri@hanken.fi

References

- Björk, B-C. (2007). A model of scientific communication as a global distributed information system, *Information Research,* **12**(2), paper 307. Retrieved 11 January, 2009 from http://InformationR.net/ir/12-2/paper307.html (Archived by WebCite® at http://www.webcitation.org/5dkj1GfSJ)
- Budapest Open Access Initiative. (2002). Retrieved 11 January, 2009 from http://www.soros.org/openaccess/read.shtml (Archived by WebCite® at http://www.webcitation.org/5dkjBzCIl)
- Ellis, D. (2005). Ellis's model of information-seeking behaviour. In Karen E. Fisher, Sanda Erdelez & Lynne Mckechnie, (Eds.). *Theories of information behavior* (pp. 138-142). Medford, NJ: Information Today.
- European Commission. *Directorate General for Research*. (2006). *Study on the economic and technical evolution of the scientific publication markets in Europe. Final report - January 2006* Brussels: European

Commission. Directorate General for Research. Retrieved 11 January, 2009 from http://ec.europa.eu/research/science-society/pdf/scientific-publication-study_en.pdf (Archived by WebCite® at http://www.webcitation.org/5dkkqPKKQ)

- Hajjem, C., Harnad, S., & 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*, **28**(4), 39-47. Retrieved 11 January, 2009 from http://eprints.ecs.soton.ac.uk/11688/1/ArticleIEEE.pdf (Archived by WebCite® at http://www.webcitation.org/5dklR0gu6)
- Harnad, S., Brody, T., Valliéres, F., Carr, L., Hitchcock, S., Gingras, Y. and others. (2004). The access/impact problem and the green and gold roads to open access. *Serials Review*, **30**(4), 310-314. Retrieved 11 January, 2009 from http://eprints.ecs.soton.ac.uk/10209/1/impact.html (Archived by WebCite® at http://www.webcitation.org/5dklK7unk)
- Hedlund, T., Gustafson, T. & Björk, B.-C. (2004). The open access scientific journal: an empirical study. *Learned Publishing*, **17**(3), 199-209.
- McVeigh, M.E. (2004). *Open access journals in the ISI citation databases: analysis of impact factors and citation patterns: a citation study from Thomson Scientific*. New York, NY: Thomson Scientific. Retrieved 11 January, 2009 from http://scientific.thomson.com/media/presentrep/essayspdf/openaccesscitations2.pdf (Archived by WebCite® at http://www.webcitation.org/5dkllnR3f)
- Morris, S. (2007). Mapping the journal publishing landscape: how much do we know? *Learned Publishing*, **20**(4), 299-310.
- Regazzi, J. (2004). The shifting sands of open access publishing: a publisher's view. *Serials Review*, **30** (4), 275-280.
- SHERPA. (2008). *RoMEO: Publisher copyright policies and self-archiving.* Nottingham, UK: University of Nottingham. Retrieved 11 January, 2009 from http://www.sherpa.ac.uk/romeo/ (Archived by WebCite® at http://www.webcitation.org/5dkm5SlgB)
- Tenopir, C. & King, D. (2000). *Towards electronic journals: realities for scientists, librarians and publishers.* Washington, DC: Special Libraries Association.
- U.K. *Parliament. Select Committee on Science and Technology* (2004). *Scientific publications: free for all? Tenth Report of Session 2003-04. Volume 2. Oral and written evidence. Appendix 46. Memorandum from Reed Elsevier.* . London: Parliament. Retrieved 11 January from http://www.publications.parliament.uk/pa/cm200304/cmselect/cmsctech/399/399we57.htm (Archived by WebCite® at http://www.webcitation.org/5dkjyOBj3)

How to cite this paper

Björk, B-C., Roos, A. & Lauri, M. (2009). "Scientific journal publishing: yearly volume and open access availability" *Information Research*, 14(1) paper 391. [Available from 12 January, 2009 at http://InformationR.net/ir/14-1/paper391.html]

Find other papers on this subject

Scholar Search Google Search Windows Live

Bookmark This Page

Last updated: 11 January, 2009

Contents | Author index | Subject index | Search | Home