

NATURE COMMUNICATIONS: CITATION ANALYSIS



Background

Nature Communications is a multidisciplinary journal, publishing high-quality research from all areas of the natural sciences. It was established in 2010. Authors can publish either through the traditional subscribed access route or make their paper open access (OA) through payment of an article-processing charge (APC). A large number of authors have chosen the OA option. This report seeks to answer the questions:

- Are OA articles cited more frequently than non-OA articles?
- Are OA articles cited sooner than non-OA articles?
- Are there differences by subject area in patterns of citation for OA and non-OA articles?
- Are there differences in web-based activity (HTML views and PDF downloads) for OA and non-OA articles?

The analysis is based on data provided by NPG on the numbers of articles published each year as they were assigned to four subject areas, and of citations to those articles as recorded in Web of Knowledge. This report is intended to be an exploratory white paper, analysing machine-generated data on a bounded set of articles. The report seeks only to present the raw normalized data, and does not investigate the reasons for the observed differences. The reasons may be complex and due to a number of factors, and more research is needed to understand them fully, but none of these factors were part of the original scope of our report.

Data cleaning and validation

The data we received covered 2,878 articles published between April 2010 and December 2013, with citation data for all articles from publication up to April 2014.

We removed 4 articles because Web of Knowledge was not able to provide us with any data about the number of citations they had received. On reflection, we then removed 866 articles which were published on or after 1 July 2013. This was to ensure that articles with little time to gather citations did not skew the data for 2013.

Data for the final question on web-based activity are a sub-set of the original dataset, comprising only the 722 articles published in the first six months of 2013.

Descriptive statistics

Figure 1 shows the total number of articles published between April 2010 and 30 June 2013 in the four subject areas of biological sciences, chemistry, earth sciences, and physics. It shows that more than half the published articles were in the biological sciences, falling to under 4% in earth sciences. The proportion published on OA terms varied from 41.0% in biological sciences to 30.4% in chemistry.

FIGURE 1: OA BY SUBJECT

			SUBJECT				Total
			Biological sciences	Chemistry	Earth sciences	Physics	
Open Access?	Subscription	Count	623	156	51	423	1253
	Count as a % within subject		59.0%	69.6%	68.9%	64.7%	62.4%
	OA	Count	433	68	23	231	755
	Count as a % within subject		41.0%	30.4%	31.1%	35.3%	37.6%
Total		Count	1056	224	74	654	2008
		Count as a % across all subjects	52.6%	11.2%	3.7%	32.7%	100%

Figure 2 (overleaf) shows for all subjects the numbers of articles published on OA terms and on a subscription basis in each of the years 2010-2012, and in the first six months of 2013. It indicates that the proportion of articles published OA has been falling since 2011.

FIGURE 2: OA BY PUBLICATION YEAR

			Publication Year				Total
			2010	2011	2012	2013 (to end of June)	
Open Access?	Subscription	Count	80	245	435	493	1253
		% within publication year	56.3%	55.3%	62.1%	68.3%	62.4%
	OA	Count	62	198	266	229	755
		% within publication year	43.7%	44.7%	37.9%	31.7%	37.6%
Total		Count	142	443	701	722	2008

Figure 3 shows the pattern for the total number of articles published – whether on subscription or on OA terms - each year by broken down by subject. It indicates that there has been no obvious change in the proportions of articles assigned to the four subject groups, although there may be some small sign of an increase in the proportion for biological sciences, matched by a fall for physics, in the first half of 2013.

FIGURE 3: PUBLICATION YEAR BY SUBJECT

			SUBJECT				Total
			Biological sciences	Chemistry	Earth sciences	Physics	
Publication Year	2010	Count	73	15	7	47	142
		% within publication year	51.4%	10.6%	4.9%	33.1%	100%
	2011	Count	232	46	18	147	443
		% within publication year	52.4%	10.4%	4.1%	33.2%	100%
	2012	Count	359	78	19	245	701
		% within publication year	51.2%	11.1%	2.7%	35.0%	100%
	2013 (to end of June)	Count	392	85	30	215	722
		% within publication year	54.3%	11.8%	4.2%	29.8%	100%
Total		Count	1056	224	74	654	2008
		% within publication year	52.6%	11.2%	3.7%	32.6%	100%

Figure 4 (overleaf) shows the numbers and proportions of articles published on OA and subscription terms by year and by subject. It indicates that while the numbers published on OA terms in biological sciences has increased, the proportion of all articles in that subject area published OA has fallen from 58.9% in 2010 to 34.2% in the first half of 2013. In other subject areas the proportion published OA has tended to fluctuate from year to year, though in chemistry and in earth sciences the size of those fluctuations may simply reflect the much smaller numbers in those subject areas. The data provided to us does not permit us to undertake an analysis which controls for possible masking effects that may affect the year-on-year and discipline-based analysis here.

FIGURE 4: OA BY PUBLICATION YEAR AND SUBJECT

		Biological Sciences		Chemistry		Earth Sciences		Physics	
Year		Subs	OA	Subs	OA	Subs	OA	Subs	OA
2010	Count	30	43	10	5	5	2	35	12
	% within subject and year of publication	41.1%	58.9%	66.7%	33.3%	71.4%	28.6%	74.5%	25.5%
2011	Count	125	107	29	17	12	6	79	68
	% within subject and year of publication	53.9%	46.1	63.0%	37.0%	66.7%	33.3%	53.7%	46.3%
2012	Count	210	149	55	23	14	5	156	89
	% within subject and year of publication	58.5%	41.5%	70.5%	29.5%	73.7%	26.3%	63.7%	36.3%
2013 (to end of June)	Count	258	134	62	23	20	10	153	62
	% within subject and year of publication	65.8%	34.2%	72.9%	27.1%	66.7%	33.3%	71.2%	28.8%

Are open access articles cited more than subscription articles, and does this vary by discipline?

It is important to note in presenting any analysis of citations of OA and subscription articles that we cannot control for all possible factors, including the possibility that many of the articles published on a subscription basis may have been deposited and made accessible via institutional or subject-based repositories, or pre-print services such as ArXiv. Nor has the data provided to us enabled us to control for factors such as the number of authors of each paper, or their geographical location, or the possibility that authors may choose to make their 'best' papers OA. *Nature Communications*, launched in 2010, is a young journal which is still developing its brand and its author and subscriber base. Factors such as these may well have had an impact on citation scores.

With that important qualification in mind, we nevertheless show in Figures 5 and 6 respectively an analysis of patterns in the numbers of citations received by all 2008 papers in the sample, broken down by subject and by OA/subscription status; and of the statistical significance of any differences by subject and by status. Figure 5 shows the maximum, median and minimum numbers of citations, together with 1st and 3rd quartiles. Figure 6 shows the results of a two-tailed Wilcoxon rank-sum test, with W, z, p and r values. We used SPSS to undertake the analysis, and calculated the effect size (r) following Field¹.

This analysis shows that OA content is cited more than subscription content when looking at the journal as a whole ($p < .001$; effect size = -.16) though the effect is small². The same applies to all subject areas except chemistry, where any OA/subscription effect on citations is not statistically significant.

FIGURE 5: FREQUENCY CHARTS FOR OA AND SUBSCRIPTION CITATIONS BY DISCIPLINE

		All Papers		Biological Sciences		Physics		Chemistry		Earth Sciences	
		Subs	OA	Subs	OA	Subs	OA	Subs	OA	Subs	OA
Max 3rd quartile Median 1st quartile Min N	Max	248	256	176	137	166	256	248	135	34	135
	3rd quartile	15	21	11	19	21	25	25	33	6	17
	Median	7	11	5	9	9	14	11	15	2	6
	1st quartile	3	4	2	3	4	6	4	5	0	2
	Min	0	0	0	0	0	0	0	0	0	0
	N	1253	755	623	433	423	231	156	68	51	23

¹ For more detail on how SPSS performs the Wilcoxon rank-sum test, including why it always produces negative z values, see <http://www-01.ibm.com/support/docview.wss?uid=swg21479010>. Effect size (r) calculated following Field, A P (2009) 'Discovering Statistics using SPSS', 3rd edition. London, SAGE Publications Ltd (p.550).

² See Cohen, 1988, Statistical power analysis for the behavioural sciences, 2nd edition. New York, Academic Press

FIGURE 6: STATISTICAL SIGNIFICANCE OF DIFFERENCE BETWEEN OA AND SUBSCRIPTION CITATIONS BY DISCIPLINE

Discipline	W	z	p	r
All disciplines	1167209.50	-7.273	.000	-.16
Biological Sciences	296828.00	-6.663	.000	-.21
Chemistry	17067.50	-1.083	.279	-.07
Earth sciences	1683.00	-2.704	.007	-.31
Physics	129424.50	-3.947	.000	-.15

Our analysis also showed that 120 of the articles published between 2010 and June 2013 had not been cited at all by April 2014. Of these:

- 86/120 were published on subscription terms (thus 7% of all subscription articles had not been cited)
- 34/120 were published on OA terms (thus 5% of all OA articles had not been cited).

At the other end of the citation spectrum, of the top 101 cited articles:

- 58/101 were published on subscription terms (5% of all subscription articles published). And it is worth noting that the highest cited paper in biological sciences was a subscription article.
- 43/101 were published on open access terms (6% of all open access articles published).

We also analysed Altmetric data for OA and subscription articles, but found no statistically significant differences either within the full dataset or when broken down by discipline.

Are open access articles cited more frequently in the short term and do citations for open access and subscription articles even out over the long term?

To answer this question we have looked at citations to older and newer articles and compared OA and subscription articles within each year to see whether a citation advantage exists in either or both groups and whether this varies over time. We are aware of the limitations of this method, since we cannot control for the possibility that other factors, such as increased awareness of the journal on the part of readers or greater prestige of the journal leading to 'better' contributions, are affecting the difference between subscription and OA citations, rather than the simple passage of time.

A more fruitful and accurate analysis would have examined citations one and two years after publication for all articles where this is possible (i.e. those published up to April 2012). By comparing the one-year and two-year citation differences for the same sets of OA and subscription articles, we could have been more confident that our analysis identifies differences over time, rather than those influenced by other factors. This was not possible with the data provided.

Figures 7 and 8 show respectively an analysis of patterns in the numbers of citations received by papers published each year by OA/subscription status; and of the statistical significance of any differences by status. Again, Figure 7 shows the maximum, median and minimum numbers of citations, together with 1st and 3rd quartiles; and Figure 8 shows the results of a two-tailed Wilcoxon rank-sum test, with W, z, p and r values.

FIGURE 7: FREQUENCY CHARTS FOR CITATIONS TO OA AND SUBSCRIPTION PAPERS BY YEAR OF PUBLICATION

	All papers		2010		2011		2012		2013 (to end of June)	
	Subs	OA	Subs	OA	Subs	OA	Subs	OA	Subs	OA
Max	248	256	248	135	176	137	116	125	50	256
3rd quartile	15	21	41.5	38.25	28	33	15	20.25	6	7
Median	7	11	26	22.5	15	20	8	12	3	3
1st quartile	3	4	14.5	11	9	13	4	6	1	1
Min	0	0	4	4	0	2	0	0	0	0
N	1253	755	80	62	245	198	435	266	493	229

FIGURE 8: STATISTICAL SIGNIFICANCE OF DIFFERENCE BETWEEN CITATIONS TO OA AND SUBSCRIPTION PAPERS BY YEAR OF PUBLICATION

Year of publication	W	z	p	r
2010	4141.00	-1.201	.230	-.10
2011	49430.50	-3.703	.000	-.18
2012	140604.00	-4.648	.000	-.18
2013 (to end of June)	173971.50	-1.639	.101	-0.06

This analysis shows that for articles published in 2011 and 2012, OA papers were cited more frequently than subscription papers, though again the effect ($p < .001$ and effect size $-.18$) was small. The same was not true for 2010 and 2013, where no statistically significant difference was observed.

For 2013 in particular the overall number of citations at each quartile was much lower for both OA and subscription content than in previous years, suggesting that these articles have not yet reached their full citation potential. This seems to carry a number of implications. First, that the OA citation advantage may last a relatively long time (articles published in 2011 would be 2-3 years old when citation data collection occurred). Second, it suggests that any OA citation advantage is not obvious at a particularly early stage (articles published in the first half of 2013 would have been up to 15 months old at the time of citation data collection). Finally, it suggests that the citation half-life of articles may be quite long.

The results for 2010 may indicate that after a considerable period of time (3-4 years) the OA citation advantage tails off. But it may also reflect citation patterns to a brand new journal and the choices made by authors about which articles to publish under which business model in that journal.

Using the one year/ two year approach to data collection outlined above may help to test all of these hypotheses further³.

Finally, Figures 9 and 10 show the pattern of Altmetric scores for OA and subscription articles by year, and the statistical significance of any differences. In this case, the differences between OA and subscription articles were statistically significant for the years 2010 and 2012 (though again the effects were small), but not for 2011 and 2013. We find it difficult to identify any theoretical justification for the differences, and we cannot exclude simple random fluctuations in the data.

FIGURE 9: FREQUENCY CHART FOR OA AND SUBSCRIPTION ALTMETRIC DATA BY YEAR OF PUBLICATION

	2010		2011		2012		2013 (to end of June)	
	Subs	OA	Subs	OA	Subs	OA	Subs	OA
Max	34.5	39.6	52.6	97.8	180.2	197.3	683.8	395.0
3rd quartile	0	6.2	5.1	5.3	9.5	11.8	15.5	13.9
Median	0	0	0	1	1.5	2.6	1.5	2.3
1st quartile	0	0	0	0	0.3	0.5	0.3	0.3
Min	0	0	0	0	0	0	0	0
N	80	62	245	198	435	266	493	229

FIGURE 10: STATISTICAL SIGNIFICANCE OF DIFFERENCE BETWEEN OA AND SUBSCRIPTION ALTMETRIC DATA BY YEAR OF PUBLICATION

Year of publication	W	z	p	r
2010	5250.50	-2.323	.020	-.19
2011	53096.00	-.992	.321	-.05
2012	147108.00	-2.151	.031	-.08
2013 (to end of June)	176428.00	-.692	.489	-.02

³We should, of course, be happy to undertake such an analysis.

ARE OPEN ACCESS ARTICLES VIEWED OR DOWNLOADED ONLINE MORE FREQUENTLY THAN THOSE PUBLISHED UNDER THE SUBSCRIPTION MODEL?

Additional data were supplied to enable us to undertake analysis of web-based usage for 722 papers published in the first six months of 2013. In total, 493 articles were published through subscribed access, and 229 were published via open access. Data were supplied for the total number of HTML page views at 30 and 180 days post-publication, and the total number of PDF downloads at 30 and 180 days post-publication.

Figure 11 shows descriptive statistics for the usage data

FIGURE 11: FREQUENCY CHART FOR ONLINE ACTIVITY DATA FOR OA AND SUBSCRIPTION ARTICLES AT 30 AND 180 DAYS POST-PUBLICATION

	30 day HTML views		180 day HTML views		30 day PDF downloads		180 day PDF downloads	
	Subs	OA	Subs	OA	Subs	OA	Subs	OA
Max	9418	34536	11061	39193	4775	16673	12215	17438
3rd quartile	704	1838	1231	3121	336	853	630	1481
Median	456	1142	804	2051	216	499	399	904
1st quartile	312	845	563	1444	142	331	259	615
Min	40	181	113	586	21	6	56	10
N	493	229	493	229	493	229	493	229

Data were analysed using the Wilcoxon rank-sum test and results are presented in Figure 12.

FIGURE 12: STATISTICAL SIGNIFICANCE OF DIFFERENCES BETWEEN OA AND SUBSCRIPTION ARTICLES FOR ONLINE ACTIVITY AT 30 AND 180 DAYS POST-PUBLICATION

Measure		W	z	p	r
HTML Views	30 days	136993.5	-15.807	.000	-.588
	180 days	135165.5	-16.508	.000	-.614
PDF Downloads	30 days	142149.5	-13.830	.000	-.515
	180 days	143072.0	-13.476	.000	-.502

For each measure, the open access articles received significantly more usage than those published under a subscription model. In each case, the effect size was large. This suggests that papers have increased visibility and discoverability online, although of course we can say little about the subsequent use of papers that are viewed or downloaded by users.

Conclusions

Overall, articles published OA appear to show a higher number of citations, though the effect is small, and the data provided does not allow us to control for possible confounding effects such as the posting of articles in repositories, the number and location of authors, and the possibility that authors are selecting their 'best' papers to publish on OA terms. Similarly, any effect of OA on the timing of citations appears to be small, and we have not been able to control for possible changes such as increased awareness of the journal on the part of both readers and authors. But although the impact on citations is small, the impact of open access publication on HTML views and PDF downloads is large and significant, suggesting increased visibility for the open access papers.