Bibliometric study on Dutch Open Access published output 2000-2012/2013

Preliminary research report to the Dutch Ministry of Science, Culture & Education (OC&W)

January 2015

Center for Science & Technology Studies (CWTS)
Leeuwen@CWTS.NL

Contents

1. Introduction

This preliminary report is part of a larger study on Open Access publishing by Dutch scientists and scholars over roughly the last decade. The wider study includes more data sources, which are not yet all sorted out properly, and are in this stage of the research not sufficiently well analyzed to include and draw firm conclusions on this data. In this text we will primarily focus on the way Open Access (OA from now on) publications are represented in the Web of Science database. We have collected data for this analysis in two different ways, which leads to different perspectives on OA publishing in the Netherlands.
We focus on the output of three smaller scientific nations in Europe, next to the Netherlands we focus on Denmark and Switzerland, as these countries do contest the scientific runner up positions globally after the USA, and are more or less of comparable volume in economic terms.

2. Data collection

In the study we make use of data from various sources, the first to be mentioned here is the Web of Science (WoS) database. We use the internet version, available to most Dutch researchers through their desktops on their PC. However, within CWTS we have a bibliometric variation of that a database, a tailor-made database that allows for bibliometric analysis, based upon the state-of-the-art bibliometric techniques and indicators. However, in this version, the functionality to search for OA output is not yet available. Finally, we make use of the journals and the publications therein listed on the Directory of Open Access Journals (DOAJ). From this data source, we will further focus on the digital object identifier's (doi's), while leaving out other important elements (such as the license types also out of the analysis, as this information is also unclearly defined in that database, as well as unclearly attached to the publications).

Method I: The first way of data collection from WoS starts form the desktop interface of the WoS database. The functionality to collect this information is not yet available in the in-house WoS database at CWTS, so therefore we had to collect these data from the internet version directly. This approach followed these steps:
Collect the output of one of the selected countries for a particular year;
Within that set, further distinguish the OA part of that selected output;
Download these publications from the WoS database (including the so-called UT-code, a unique identifier within WoS, that allowed for further linking to the CWTS WoS database);

Select within the CWTS database the output for the three countries;
Match the selected output form the Internet version of the WoS with the in-house CWTS version;
Leading to two sets within the CWTS database, an OA formatted set of publications, and a non OA formatted set of publications.
These steps were taken for all three countries, collecting publications from 2000-2013.
The definition of how the publications were defined as OA is based upon the following statement on the WoS database' website: "The Thomson Reuters Links Open Access Journal Title List includes free journal content that are available for linking from the Web of Science.".

Method II: The second method started from the Directory of Open Access Journals (DOAJ) list. This list contains journals that have the Gold Open Access business model. CWTS has downloaded the complete list, and all publications published in the journals on the DOAJ list. By making use of this dataset, we could come to a second approach to the OA output of the three countries. Therefore, we followed the next steps:
First select within the CWTS database the output for the three countries;
Collect their Digital Object Identifiers (doi);
match these with the doi's of the publications downloaded from the DOAJ list;
Leading to two sets within the CWTS database, an OA formatted set of publications, and a non OA formatted set of publications.
Please note that we focused on articles, letters and reviews only, excluding other types of documents, such as editorials, meeting abstracts, book reviews, etc. The choice for these types is based upon the importance of these three types in communicating scientific findings among peers, and their relative homogeneity within the system.

3. Methodology

In the study we present a number of indicators. In cases we present numbers of publications, this is indicated with a P. In case citation data are presented, we use the non-normalized indicators MCS (Mean Citation Score) and TCS (Total Citation Score). Furthermore, we use MNCS (Mean Normalized Citation Score), as well as the MNJS, the field normalized journal impact indicator, to indicate the normalized impact scores in the study ((Waltman et al, 2011a, and Waltman et al, 2011b). While the output indicator can be used for the various electronic systems we use in the study, and P can relate to various documents types analyzed, the citation impact indicators are used only within the context of the WoS

database. Please note that in case of the impact indicators, the length of the citation window is one year longer as expressed by the presented year block (so in case of the last block, 2009-2012, the citation impact is measured up until 2013, currently the last year fully covered in the CWTS WoS database). Other indicators that we use are the PP TOP 10% (the proportion of publications from a set that belongs to the top 10% most highly cited documents in the field(s)). Other indicators used are the percentage not cited publications (PP NOT CITED %) and the percentage self-citations (PP SELF CITS %).

4. Results

First we present the results form Method I, described above. The output numbers of the three countries according to the methodology I are found in tabular form in Table 1. The analysis covers the period 2000 up until 2012 for publication data, and up until 2013 for citation impact data. In this analysis we have combined several years together, in order to create more solid and stable trend lines, as we mostly interested in the trends than in variation from year to year. Table 1 below contains the output numbers from 2000 onwards, for the three countries, and the two separate parts of the output, distinguished by openness.

Table 1. Output (P) of Denmark, the Netherlands, and Switzerland, distinguishing OA and non OA output, 2000-2012

	NL			DK			CH		
	NL Ex OA	NL OA	Share OA	DK Ex OA	DK OA	Share OA	CH Ex OA	CH OA	Share OA
2000 - 2003	75607	712	1%	30616	452	1%	53283	995	2%
2001 - 2004	78087	858	1%	31262	557	2%	54793	1220	2%
2002 - 2005	81849	1180	1%	31972	728	2%	56982	1836	3%
2003 - 2006	85386	1663	2%	33024	949	3%	60319	2217	4%
2004 - 2007	88745	2349	3%	34082	1244	4%	63205	2790	4%
2005 - 2008	92349	3265	4%	35273	1631	5%	65920	3517	5%
2006 - 2009	96278	4269	4%	36672	1997	5%	69518	3912	6%
2007 - 2010	101270	5587	6%	38726	2554	7%	72687	4981	7%
2008 - 2011	106560	7299	7%	41417	3264	8%	76658	6354	8%
2009 - 2012	111990	9504	8%	44264	4420	10%	80786	7990	10%

The data presented in Table 1 clearly show that OA publishing is becoming increasingly important, in all three selected countries. The Netherlands is relatively lagging somewhat behind compared to Denmark and Switzerland, albeit with only a small part of the total output, perhaps influenced by the larger volume of the total output of the Netherlands as well.

In Figure 1, we have distinguished between the Open Access format output of the three countries, and we have taken that out of the total national output on these three countries (indicated by the 'Ex OA' label to the country names). What we observe are increasing trends for the parts of the output not published in OA format, which is also visible for the OA format of the output of these three countries, and as shown above in Table 1, increases somewhat faster for Denmark and Switzerland as compared to the Netherlands.

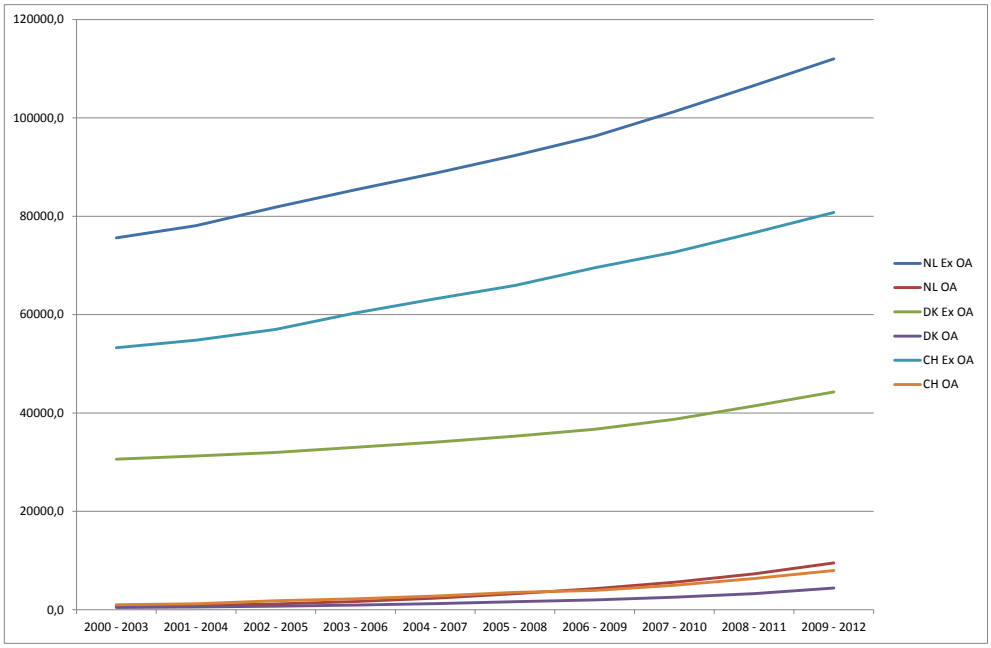

Figure 1. Output development (P) of Denmark, the Netherlands, and Switzerland, 2000-2012/2013.

In Table 2, we present the citation impact scores as represented by the MNCS indicator, the field normalized impact of the outputs of the three countries, again separated by the two types of publication output: Open Access and non-Open Access publications.

Table 2. Citation impact (MNCS) of Denmark, the Netherlands, and Switzerland, distinguishing OA and non OA output, 2000-2012

	NL		DK		CH	
	NL Ex OA	NL OA	DK Ex OA	DK OA	CH Ex OA	CH OA
2000 – 2003	1,29	0,99	1,30	1,03	1,37	1,11
2001 - 2004	1,30	0,95	1,29	1,31	1,35	1,21
2002 - 2005	1,30	0,99	1,29	1,39	1,36	1,36
2003 - 2006	1,31	1,07	1,31	1,34	1,36	1,46
2004 - 2007	1,30	1,12	1,31	1,30	1,38	1,47
2005 - 2008	1,31	1,13	1,32	1,30	1,39	1,48
2006 - 2009	1,35	1,15	1,34	1,26	1,39	1,39
2007 - 2010	1,38	1,17	1,37	1,26	1,42	1,37

2008 - 2011	1,40	1,18	1,40	1,25	1,46	1,36
2009 - 2012	1,44	1,18	1,44	1,18	1,50	1,33

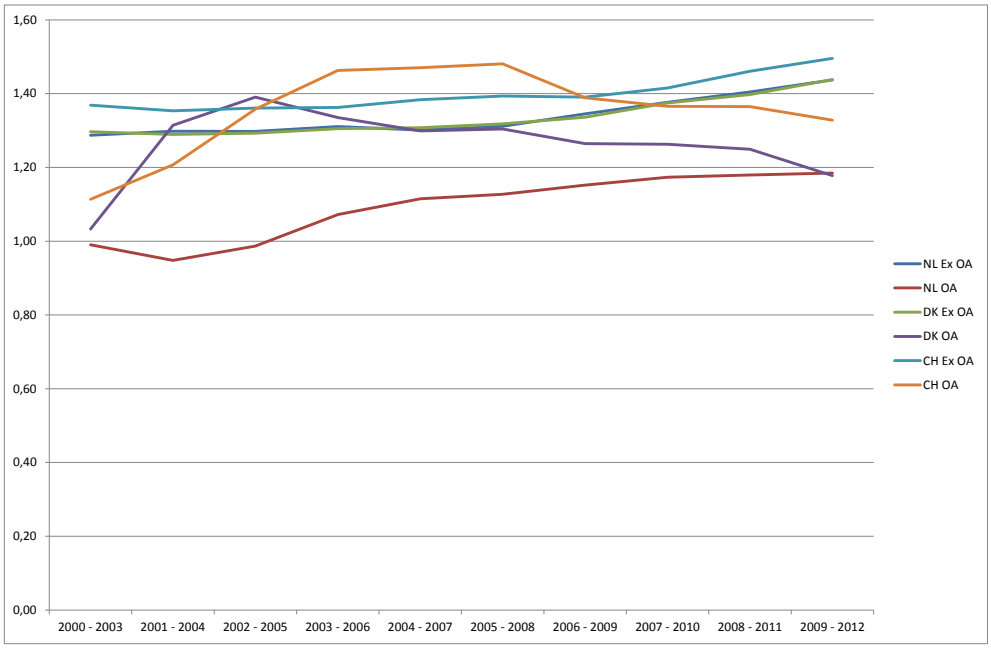

Figure 2. Impact development (MNCS) of Denmark, the Netherlands, and Switzerland, 2000-2012/2013.

Figure 2 shows that for all three countries the non OA part of the output has an impact well above worldwide average impact level, with Switzerland topping the other two countries, which have a nearly equal field normalized impact score. Interesting is here that the impact of the OA part of the national outputs of Denmark and Switzerland were initially well above worldwide average impact level, and even in the case of Switzerland, for a longer period, as the OA format published output is lower on MNCS only from 2007-2010/2011 onwards. In case of Denmark, this drop started somewhat earlier, while in the case of the Netherlands, the OA output never got an impact higher as compared to the non OA format output. Another interesting phenomenon is the increase of the gap between the impact of OA and non OA output, particularly for Switzerland and Denmark, where we observe a clear seizure of the impact of OA format output, as compared to their non OA formatted output, and to a lesser extent for the Netherlands, where the two impact lines are more slowly diverging. If we shift our focus towards the journal impact analysis (see Table 3 and Figure 3), for which we use the indicator MNJS, we see an even more interesting phenomenon. While the output in non OA format published journals shows a choice for journals with increasing impact scores (with here the Swiss and the Dutch topping, and nearly overlapping, and the Danish output is published in journals with a somewhat lower, but still increasingly higher impact),

the OA format published outputs end up in journals with decreasing field normalized impact scores. We even notice a diverging trend in these two clusters of trend lines: non OA format published journal tend to show increasing impact scores, while OA format published journals show decreasing impact trends. This is even more striking if we consider that we here look at three of the 'scientifically stronger' nations, as far as can be measured with bibliometric instruments.

Table 3. Journal-to-field citation impact (MNJS) of Denmark, the Netherlands, and Switzerland, distinguishing OA and non OA output, 2000-2012

	NL		DK		CH	
	NL Ex OA	NL OA	DK Ex OA	DK OA	CH Ex OA	CH OA
2000 - 2003	1,18	0,95	1,15	0,84	1,19	1,06
2001 - 2004	1,19	0,97	1,16	1,02	1,20	1,03
2002 - 2005	1,19	1,00	1,16	1,08	1,20	1,19
2003 - 2006	1,20	1,06	1,16	1,11	1,20	1,20
2004 - 2007	1,22	1,09	1,18	1,12	1,22	1,11
2005 - 2008	1,24	1,09	1,20	1,10	1,24	1,14
2006 - 2009	1,26	1,11	1,22	1,07	1,26	1,11
2007 - 2010	1,29	1,11	1,25	1,06	1,29	1,11
2008 - 2011	1,30	1,10	1,26	1,05	1,31	1,11
2009 - 2012	1,32	1,09	1,28	1,00	1,33	1,09

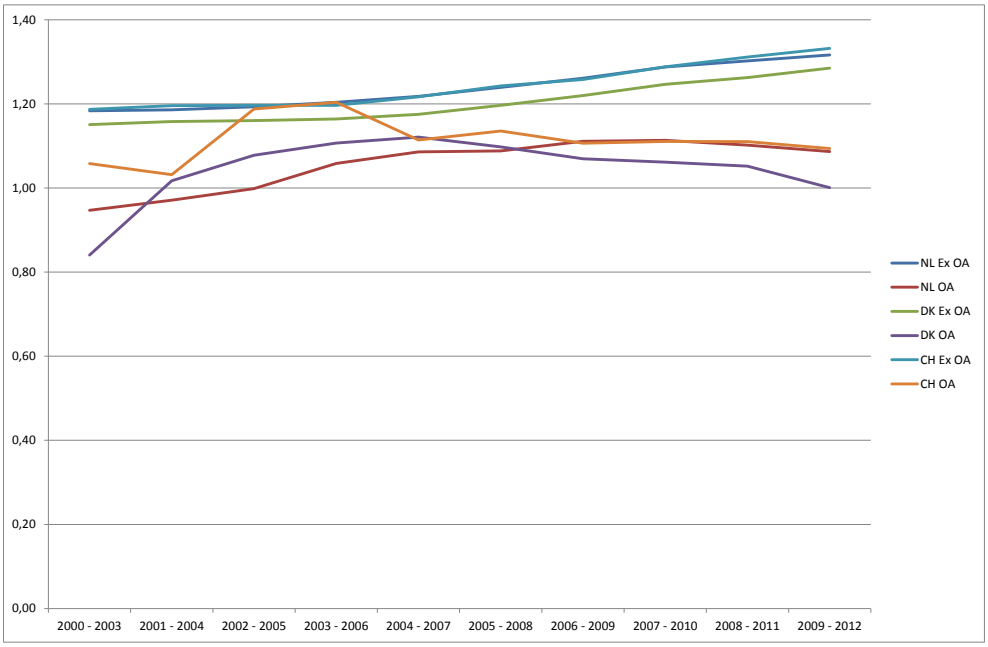

Figure 3: Journal impact development (MNJS) of Denmark, the Netherlands, and Switzerland, 2000-2012/2013

Here we start with the results form methodology II. The results of the output analysis are shown in Table 4, which again covers a similar distinction between OA and non OA format output, but now according to the definition described above under Method II. We combined the DOI's of journals on the DOAJ list with the DOI's available in the WoS. From total set of 787.611 DOI's in the DOAJ list, we matched 226.641 publications in WoS on the basis of available DOI's. The reason for this seemingly low recall is twofold. In the first place, not all journals covered by the DOAJ list are processed for the WoS database, and secondly, not all publications in journals covered in WoS do contain doi's, which means that for some journals that are both covered in the DOAJ list as well as in WoS, a match is impossible, particularly this will be the case for the earlier years in the analysis. Like the first methodology we followed, we isolated the OA format published output from the Netherlands, Denmark, and Switzerland from the total set of publications for the three countries under study.

If we now again focus on the three countries in our international comparison, we get the following impressions. First of all, we observe that the overlap between the DOAJ list/WoS combinations with Dutch/Danish/Swiss publications in WoS is much smaller as compared to the previous analysis on Dutch/Danish/Swiss output in OA format, which is highly likely the result of the missing doi's in the WoS database. This is becoming clear when we compare the results of Table 1 with those presented in Table 4. Here we find much lower shares of OA output as compared to the overall output of the three countries. This is further underlined by Figure 4, in which the OA format output of the three countries is on the low end of the graph, while we observe simultaneously a strong increase in the output of the non OA format output of the three countries.

Table 4. Output (P) of Denmark, the Netherlands, and Switzerland, distinguishing OA and non OA output (based on doi-matching), 2000-2012

	NL			DK			CH		
	NL Ex OA	NL OA	Share OA	DK Ex OA	DK OA	Share OA	CH Ex OA	CH OA	Share OA
2000 - 2003	75607	10	0%	30616	4	0%	53283	2	0%
2001 - 2004	78087	35	0%	31262	25	0%	54793	30	0%
2002 - 2005	81849	136	0%	31972	83	0%	56982	97	0%
2003 - 2006	85386	344	0%	33024	170	1%	60319	232	0%
2004 - 2007	88745	648	1%	34082	312	1%	63205	420	1%
2005 - 2008	92349	1068	1%	35273	486	1%	65920	690	1%
2006 - 2009	96278	1531	2%	36672	664	2%	69518	972	1%
2007 - 2010	101270	2207	2%	38726	924	2%	72687	1461	2%
2008 - 2011	106560	3036	3%	41417	1231	3%	76658	2062	3%
2009 - 2012	111990	3896	3%	44264	1595	4%	80786	2608	3%

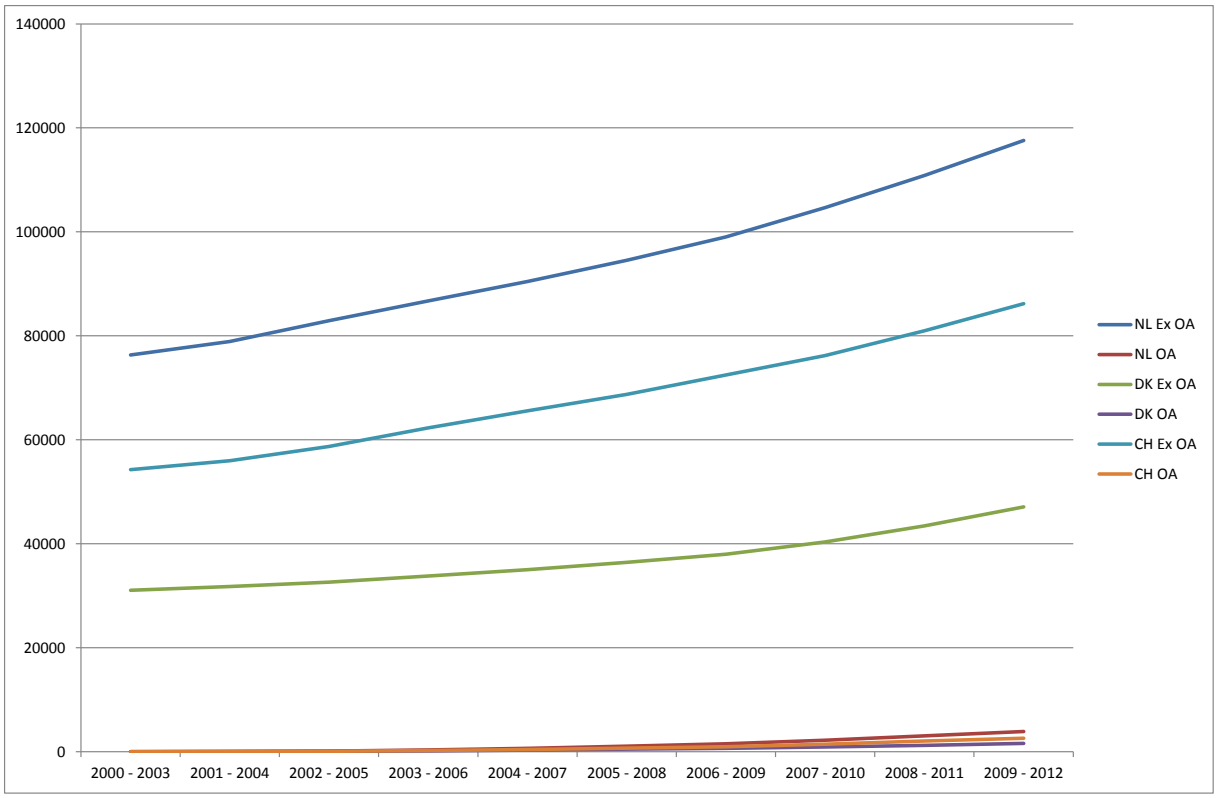

Figure 4. Output development (P) of Denmark, the Netherlands, and Switzerland, based on matching of doi's, 2000-2012/2013.

In Table 5, we present the impact scores of the three countries, again distinguishing OA format output and non OA format output. Here we again observe lower impact scores for the OA format output of the three countries, except for the starting block of the analysis (please note that the output numbers are very low in this part of the analysis for the Netherlands and Denmark, respectively 10 and 4 papers). From the second year block onwards, we observe increasing trends in the impact of the OA format of the three countries, although we must stress that this is also the case for the non OA format output of the three countries.

Table 5. Citation impact (MNCS) of Denmark, the Netherlands, and Switzerland, distinguishing OA and non OA output (based on doi-matching), 2000-2012

	NL		DK		CH	
	NL ex OA	NL OA	DK ex OA	DK OA	CH ex OA	CH OA
2000 - 2003	1,28	1,65	1,29	1,32	1,36	
2001 - 2004	1,29	0,87	1,29	0,91	1,35	1,03
2002 - 2005	1,29	0,87	1,30	0,98	1,36	1,18
2003 - 2006	1,31	0,87	1,31	0,78	1,37	0,95
2004 - 2007	1,30	0,75	1,31	0,72	1,39	0,96
2005 - 2008	1,31	0,83	1,32	0,86	1,40	0,91
2006 - 2009	1,35	0,85	1,34	0,89	1,40	0,92

2007 - 2010	1,38	0,90	1,38	0,96	1,42	0,97
2008 - 2011	1,40	0,97	1,40	1,00	1,46	1,07
2009 - 2012	1,43	1,03	1,43	0,96	1,49	1,06

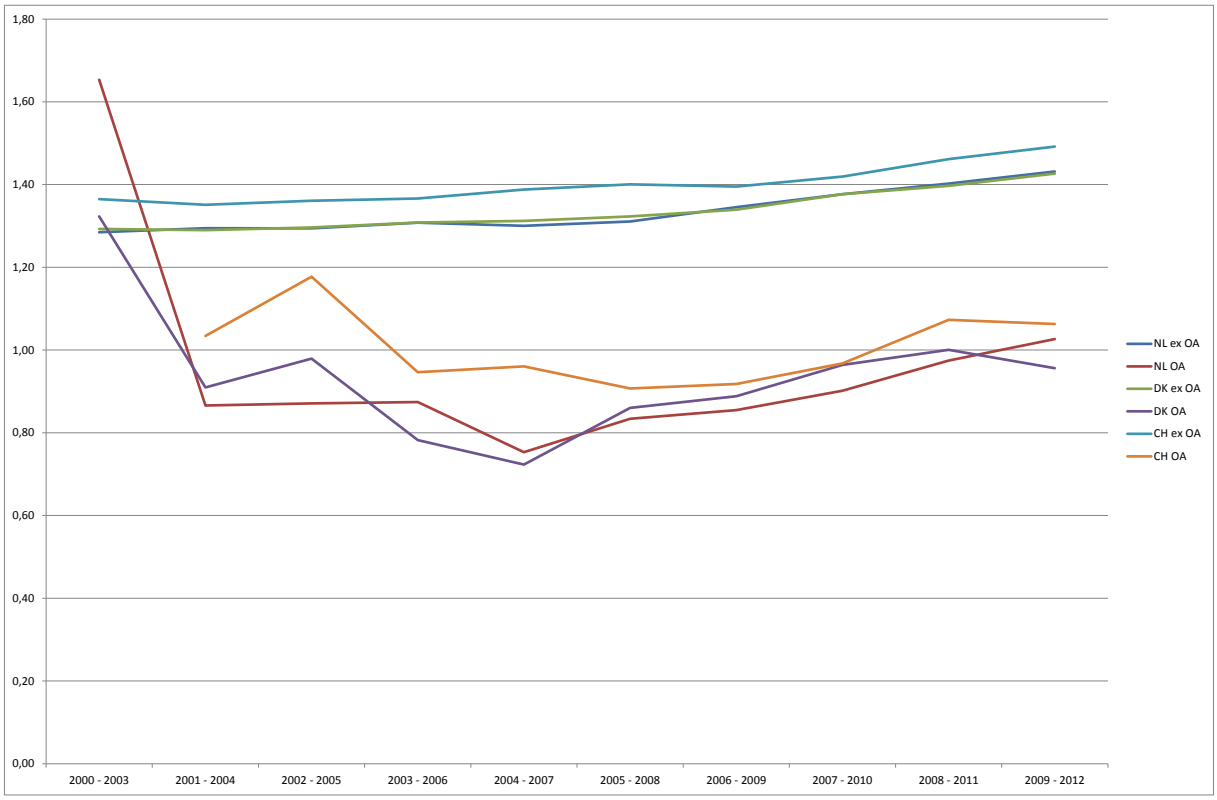

Figure 5. Impact development (MNCS) of Denmark, the Netherlands, and Switzerland, based on matching of doi's, 2000-2012/2013.

Figure 5 clearly shows this stable development of both sets of publications in time, whereby the impact scores are on both sets increasing, although the 'difference' remains more or less the same between the two sets of scores, OA and non OA format output based.

In Table 6 we present the outcomes of the analysis on the journal impact scores, based upon methodology II. Here we observe, similar to the previous outcomes, fluctuations in the initials years of the analysis for the OA format output, followed by a more stable situation from 2005-2008 onwards.

Table 6. Journal-to-field citation impact (MNJS) of Denmark, the Netherlands, and Switzerland, distinguishing OA and non OA output (based on doi-matching), 2000-2012

	NL		DK		CH	
	NL ex OA	NL OA	DKex OA	DK OA	CH ex OA	CH OA
2000 - 2003	1,18	0,54	1,15	1,28	1,19	0,24
2001 - 2004	1,18	0,84	1,16	0,92	1,19	1,22

2002 - 2005	1,19	0,77	1,16	0,84	1,20	1,00
2003 - 2006	1,20	0,84	1,16	0,79	1,20	0,90
2004 - 2007	1,22	0,86	1,18	0,83	1,22	0,88
2005 - 2008	1,24	0,88	1,20	0,86	1,24	0,86
2006 - 2009	1,26	0,90	1,22	0,87	1,26	0,87
2007 - 2010	1,29	0,94	1,24	0,91	1,29	0,91
2008 - 2011	1,30	0,97	1,26	0,93	1,31	0,96
2009 - 2012	1,31	0,97	1,27	0,92	1,32	0,97

This findings is even more visible in the graphical representation of Table 6, as in Figure 6. Here these fluctuation are very well visible, as well as the stable satiation from2005 onwards.

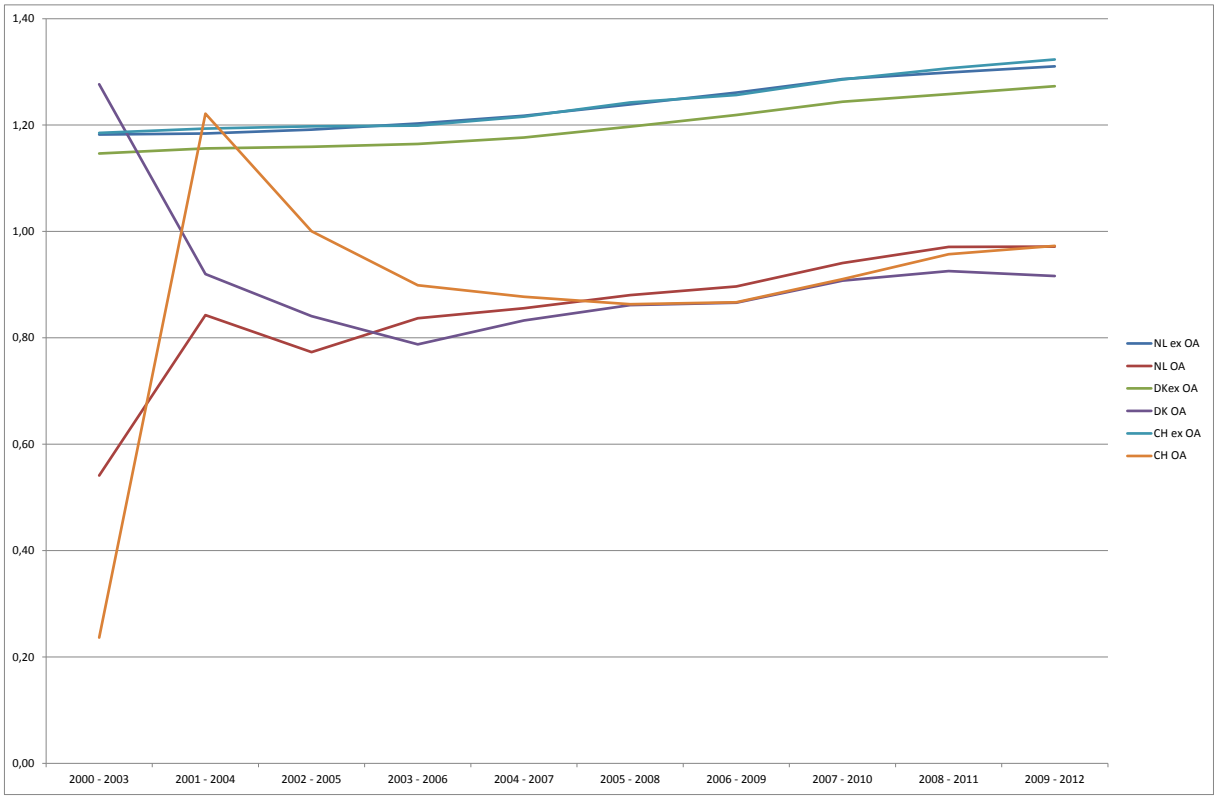

Figure 6: Journal impact development (MNJS) of Denmark, the Netherlands, and Switzerland, based on matching of doi's, 2000-2012/2013

What if nothing would happen, what if there would be no policy intervention on the publishing in Open Access sources.? Of course we can only model this in the WoS database, but if we extrapolate the current developments, that is, if we use the current growth in output in both the OA format and the non OA format output, as defined to the first methodology applied on the WoS database, we can get the following picture (figure7). What we see here is the extrapolated growth curves of these two lines of OA for the Netherlands,

and the moment in time where the total output of the Netherlands could be in Open Access. This is of course a somewhat naïve representation, as this completely ignores the fact that other, external, elements, do influence the growth in the near future (such as the way academic credits are given, and the way this works out in the ways academic careers are made.

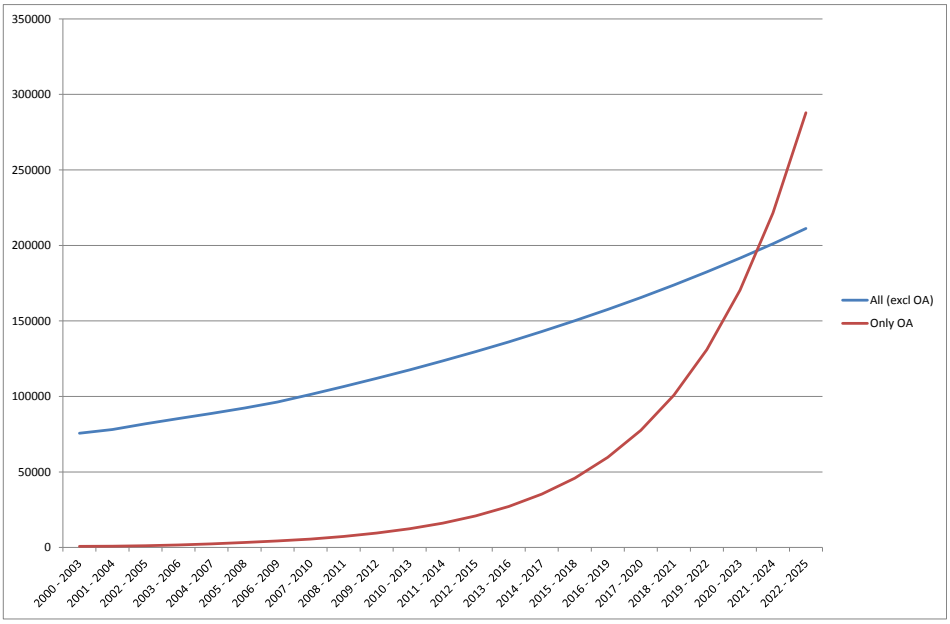

Figure 7: Simulation of the development of OA and non OA format output for the Netherlands

5. Conclusion and discussion

In this first part of the study we have learned a number of things on Open Access publishing and the way that is represented from a bibliometric perspective. As this is probably mainly due to the relative immature character of the phenomenon of Open Access, we still observe a situation in which not one clear image can be generated on the phenomenon.

So clearly a first conclusion relates to the messy situation around the various manners by which Open Access is defined, and information can be extracted from electronic databases. The two different ways Open Access can be operationalized within the world of Web of Science is a clear example of this unclear and somewhat messy situation. The fact that the Scopus database did not have the functionality to clearly defined Open Access for users of the system is another expression of the somewhat immature and unclear situation around Open Access. Further examples of this are the unclear situation around the various ways Open Access is operationalized by the publishing industry, there is no clear way of operationalizing in the larger databases of the various business models (such as Gold,

Green, and Hybrid Open Access). Yet another example relates to the various license types related to Open Access. And finally, the situation around the Narcis database, that has to be investigated much further before we can get to some conclusions on the coverage of Open Access .

With respect to our findings so far, we can say some things on the issue of the journal status within WoS database of the journals that are publishing in Open Access mode. In every environment, new developments take place to get into full swing. With this we mean that new journals have to struggle to find their position within the total 'hierarchy' of the domain, and as such also within the WoS database. There is no exception of course for OA journals, as many of these are new within the fields that they serve, or are transformations form traditional non OA journals to OA journals. This, together with the fact that in general earning academic credits is strongly related to publishing in highly visible international high impact journals, might cause these OA journals to have some trouble in getting into place within the scholarly system, and the representation of that in for example the WoS database (please note that this system mainly reflects publishing in the natural, life, and medical sciences properly, and to a much lesser extent for the social sciences, humanities, and law, see van Leeuwen, 2013).

Yet another issue to worry about is the position of peer review in the journals that are part of the expansion of the WoS database over the last couple of years, many of these in the OA segment of the database. The Institute for Scientific Information, the predecessor of the current owner of the WoS database Thomson Reuters always clearly indicated that a properly functioning peer review system within a journal was one of the conditions for a journal to be included in the system (next to other criteria, such as international focus, regularly appearance, preferably in the English language, etc.). We can now doubt whether this is still such a strong criteria, particularly given the fact that so many new journals appeared around the OA development, many of these journals trying to profit from the current developments in the market. And this is of course an issue when publishing in OA journals becomes compulsory, in any way, as is for example the case in the UK, where clear indications are given with respect to possible inclusion of output in the next REF round.

Finally, we come the conclusion of the impact scores related to the publishing in OA journals. Please note that whatever we conclude here is mainly related to the domains in which journal publishing is the dominant way of communication (so the natural, life and medical sciences). We observe for the three countries that there share in output in OA journals is lagging behind to the non OA format published material. Depending on the way you approach this question we can observe parallel developments in impact development, or even worse, a divergence of the impact related to both types of publishing, with OA on the negative side of that. This is probably due to the fact that OA is still a relative new phenomenon for many researchers and scholars, so the factor of being unknown most

certainly plays a role here, just as the other elements mentioned above. However, it is still an issue what could be the consequences for the Dutch sciences system in the somewhat longer run, particularly as OA becomes the obligatory way to communicate with your academic peer groups.

References

- van Leeuwen, T.N. Bibliometric research evaluations, Web of Science and the Social Sciences and Humanities: a problematic relationship ? Bibliometrie - Praxis und Forschung, 2013, 1-18 (http://www.bibliometrie-pf.de/article/viewFile/173/215)
- Waltman, L., N.J.van Eck, T.N. van Leeuwen, M.S. Visser, and A.F.J. van Raan, Towards a new crown indicator: Some theoretical considerations, Journal of Informetrics, 2011a, 5 (1), 37-47
- Waltman, L., N.J.van Eck, T.N. van Leeuwen, M.S. Visser, and A.F.J. van Raan, Towards a new crown indicator: An empirical analysis. Scientometrics, 2011b, 87 (3), page 467-481