

Exploring Possibilities to Use Bibliometric Data to Monitor Gold Open Access Publishing at the National Level

Thed N. van Leeuwen

CWTS, Leiden University, Wassenaarseweg 62a, Leiden, the Netherlands. E-mail: leeuwen@cwts.nl

Clifford Tatum

CWTS, Leiden University, Wassenaarseweg 62a, Leiden, the Netherlands

Paul F. Wouters 

CWTS, Leiden University, Wassenaarseweg 62a, Leiden, the Netherlands

This article¹ describes the possibilities to analyze open access (OA) publishing in the Netherlands in an international comparative way. OA publishing is now actively stimulated by Dutch science policy, similar to the United Kingdom. We conducted a bibliometric baseline measurement to assess the current situation, to be able to measure developments over time. We collected data from various sources, and for three different smaller European countries (the Netherlands, Denmark, and Switzerland). Not all of the analyses for this baseline measurement are included here. The analysis presented in this article focuses on the various ways OA can be defined using the Web of Science, limiting the analysis mainly to Gold OA. From the data we collected we can conclude that the way OA is currently registered in various electronic bibliographic databases is quite unclear, and various methods applied deliver results that are different, although the impact scores derived from the data point in the same direction.

Introduction

The implementation of policies to promote open access publications has developed a demand for accurate monitoring of the absolute and relative number of open access publications. However, open access is notoriously difficult to measure and analyses often employ random sampling techniques (Archambault et al., 2014; Björk et al., 2010). All publication records in a given sample are tested to determine the proportion of full texts that are open access publications. This method inevitably introduces many measurement errors. The implementation of new research information systems at Dutch universities and research institutes has created an opportunity to monitor the share of open access publications at the national level through coordinated metadata schemes and common registration practices. In this article, we test whether this new approach enables a more precise measurement of open access publishing.

Assessment of open access publishing is complicated by a growing diversity of what counts as open access, the copyright restrictions for when a publication can be made openly accessible, and the lack of clear and consistent identification of open access publications in bibliographic data. To examine these challenges we begin with a definition from the Budapest Open Access Initiative (BOAI 2002):

Free availability on the public internet, permitting any users to read, download, copy, distribute, print, search, or link to the full texts of these articles, crawl them for indexing, pass them as data to software, or use them for any other lawful purpose, without financial, legal, or technical barriers other than those inseparable from gaining access to the internet itself. The only constraint on reproduction and distribution,

Received March 9, 2017; revised November 6, 2017; accepted February 4, 2018

© 2018 The Authors. Journal of the Association for Information Science and Technology published by Wiley Periodicals, Inc. on behalf of Association for Information Science and Technology • Published online May 1, 2018 in Wiley Online Library (wileyonlinelibrary.com). DOI: 10.1002/asi.24029

This is an open access article under the terms of the Creative Commons Attribution-NonCommercial License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited and is not used for commercial purposes.

¹This article is an extended and revised version of the contribution to the 2015 ISSI Conference in Istanbul, Turkey. The methodology has been expanded, and consequently the text has been adjusted to this, as well as completely revised.

and the only role for copyright in this domain, should be to give authors control over the integrity of their work and the right to be properly acknowledged and cited.

This definition highlights two distinct channels of access: (i) human access to read, download, and reuse the full text of published articles and (ii) machine access to crawl, index, or analyze the content of articles. The BOAI also proposes two operational paths to access through open access journals and self-archiving in repositories, subsequently referred to as Gold open access (OA) and Green OA (Bailey, 2005). Since publication of the BOAI in 2002, definitions of OA have evolved to include a variety of formats, or types, of OA that increase the complexity of tracking progress toward increased access to published research. In Table 1, we outline common types of OA publishing found in the literature.

In addition to the broad categories of Gold and Green OA, multiple versions of a manuscript may exist due to variations in publishers' licensing agreements. These agreements typically specify how, when, and under which conditions a manuscript may be openly accessible on the web. For example, a publisher may allow Green OA through self-archiving in an institutional repository. However, publishers' copyright restrictions differ in various stages of manuscript development, thus assigning different access rights to different versions of the text. Commonly specified version types include the submitted manuscript (before peer review), the accepted manuscript (peer-reviewed but not formatted), and an exact copy of the published manuscript (Björk et al., 2013). This creates the possibility that the OA version of a manuscript is substantively different from the published version. In such instances, it is unclear whether the OA version has been validated through quality control measures such as peer review.

Delayed access creates another variation of OA. After a specified embargo period a copy of the publication may be self-archived or the publisher may completely remove access restrictions on the journal website. Embargo periods are generally specified as a delay of 6, 12, 18, or 24 months after publication, with 12 months being the most common embargo (Laakso & Björk, 2013). In the Green OA mode, it is left to authors and institutions to track and manage a variety of self-archiving policies, which in itself has been shown to be a barrier to OA (Davis & Connolly, 2007). This administrative overhead is largely absent in the case of subscription journals that convert articles to OA after a specified delay (for example, 12 months). According to Laakso and Björk, journal and article impact factors of "delayed access" journals are higher than comparable averages of subscription journals and direct (no delay) OA journals (Laakso & Björk, 2013).

A common theme in arguments for OA is that OA publishing increases citation impact. While there are conflicting reports regarding this "open access citation advantage" (OACA), heightened attention to this issue has increased our understanding about citation behavior more generally. Numerous bibliometric studies claim that OA publishing results in a significant increase in citations. In these studies

the size of advantage varies widely based on a variety of issues, such as disciplinary differences, methodological approaches, variation in how OA is defined, and difficulty in determining when an article is made openly accessible (Swan, 2010). In addition, a number of confounding factors have been shown to influence citation frequency, such as: early exposure to draft versions of a manuscript (Moed, 2007), self-selection bias, whereby an author may choose OA for only his/her best publications (Kurtz et al., 2007), the availability at multiple access points (Xia et al., 2010), and physical proximity of researchers (Lee et al., 2010).

To control for these factors, Davis et al. (2008) employ randomized controlled trial methods, whereby randomly selected articles in subscription-based journals are switched to OA. The resulting configuration is similar to hybrid OA, such that the article is made to be openly accessible and is listed among the non-OA articles on the journal's website. Davis et al. (2008) do not find a citation advantage. However, the research design used to control for confounding variables (randomized controlled trial) is most applicable to the hybrid OA. And hybrid OA publications are particularly difficult to monitor because there is not presently an agreed way of identifying them in bibliographic metadata. In other words, it would be inaccurate to apply the Davis et al. findings to all types of OA.

In a more recent study, Archambault et al. (2014) show variation in the accumulation of citations associated with different modes of OA. The authors find a citation *advantage* most prominently associated with the self-archiving mode of OA (Green OA) and a citation *disadvantage* associated with full and immediate OA journals (Gold OA). This study also establishes a general ranking of citation accumulation on the bases of OA, listed in order of most to least: Green OA, Other OA, Not OA, and Gold OA (Archambault et al., 2014: pp. 20, 24).

To address the variability of circumstances associated with OA publishing, recent studies invert the research design from top-down queries of bibliometric data sets to bottom-up testing whether a publication is an OA publication. This approach involves random sampling of a given publishing domain, harvesting full texts from the Internet, and analysis of available metadata from harvested manuscripts (Björk et al., 2010). While this approach circumvents much of the variability noted earlier, it nevertheless depends on the presence and quality of metadata.²

The objective of our analysis is to show the challenges of bibliometrically analyzing OA publications and associated impact scores. We use Web of Science (WoS) data both stand-alone and combined with article-level data extracted from journals listed in the Directory of Open Access Journals (DOAJ). To be clear, in this analysis we address the Gold OA – Journal type of OA, and not the Gold OA – Article (also known as hybrid OA; see Table 1) type of OA. As noted earlier, hybrid OA is reliably

²The potential for improved metadata practices is addressed in the Discussion section.

TABLE 1. Types of Open Access (adapted from Archambault et al., 2014 and Laakso & Björk, 2013).

Green OA	Full text (draft or published) manuscripts self-archived in a repository and/or accessible from personal, institutional, or subject websites
Gold OA, Journal Gold OA, Article	open access journals with immediate free access, some of which (for example PLoS) operate on an author pays model (Also referred to as Hybrid OA) author pays the article processing costs (APC) to make articles published in a subscription based journal that are
Delayed OA, Green	Publisher specifies an embargo period (for example 6, 12, 18, or 24 months), after which a published article may self-archived in as open access repository
Delayed OA, Journal	subscription-based journals whereby published articles are converted to open access after a specified period (for example 6, 12, 18, or 24 months)
Transient OA	Freely available on the web during a finite period (for example journal promotion); content changes in repositories and/or websites (for example updated or deleted manuscripts)
Restricted OA	Sample restrictions: access requires registration and/or membership in a group; limited use, such as read-only (not downloadable or not sharable; metadata not available for aggregation and/or analysis
Rogue OA	(also referred to as Robin Hood OA) – published manuscripts posted on websites or self-archived in repositories in conflict with licensing agreements and/or copyrights; may also contribute to transient OA

identifiable in publishers' bibliographic metadata. The analysis is focused on comparison of relative output and relative impact among three European countries of relatively comparable size and scientific production: the Netherlands, Denmark, and Switzerland. We show development over time as well as differences resulting from both approaches. It is important to note that Green OA articles are excluded from our analysis. While the Netherlands maintains a robust national repository for Green OA (NARCIS), a reliable system of identifying the self-archived state of publications within bibliometric data sets is not yet included. Our measurement of the proportion of OA and associated impact comparisons is therefore limited to Gold OA.

Data Collection

We used the WoS database in its Internet version, available to most Dutch researchers because this allows searching for OA publications. We also used the CWTS version of the WoS, a tailor-made database that consists of state-of-the-art bibliometric techniques and indicators but does not yet include the functionality to search for OA output. Finally, we made use of the journals and the publications listed in the DOAJ. From this data source, we extracted the digital object identifiers (DOIs), while leaving out other elements (such as the license types, as this information is unclearly defined as well as unclearly linked to the publications). Our study is thus limited to Gold OA of journal articles. We used three methods of data collection.

Method I: The first method of data collection starts with the desktop interface of the WoS database. This approach used the following steps:

1. Collect the output of one of the selected countries for a particular year;
2. Within that set, further distinguish the OA part of that selected output;
3. Download these publications from the WoS database (including the so-called UT-code, a unique identifier within WoS that allows for linking to the CWTS WoS database);
4. Select within the CWTS WoS database the output for the three countries;

5. Match the selected output from the Internet version of the WoS with the in-house CWTS version;
6. Create for each country two sets within the CWTS database: an OA formatted set of publications, and a non-OA formatted set of publications.
7. 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 on the following statement on the WoS database's website: "The Thomson Reuters Links Open Access Journal Title List includes free journal content that are available for linking from the Web of Science."³

Method II: The second method starts from the DOAJ. This list contains journals that have implemented the Gold OA business model. CWTS has downloaded the complete list, and all publications published in the journals on the DOAJ list. By making use of this data set, we could use a second approach to the OA output of the three countries taking the following steps:

1. First, select within the CWTS database the output for the three countries;
2. Collect their Digital Object Identifiers (DOIs);
3. Match these with the DOIs of the publications downloaded from the DOAJ list;
4. Create two sets within the CWTS database, an OA formatted set of publications, and a non-OA formatted set of publications.
5. These steps were taken for all three countries, collecting publications from 2000–2013.

Method III: The third method consists of a direct linking of DOAJ listed journals, in a double-linking process. First, by using the ISSN of the journal as the matching principle between the DOAJ list and the CWTS WoS database, and second, by using the start year of the journal in the DOAJ list (the year of becoming OA) as simulation of the publication year. There is one problem with this approach: the fact that ISSN is not a clearly distinguishing entry; some journals

³At the time the research was conducted, what is now Clarivate Analytics was then still named Thomson Reuters.

TABLE 2. Output (P) of Denmark, the Netherlands, and Switzerland, distinguishing Gold OA and all other output, 2000–2012.

	NL All Other	NL Gold OA	Share Gold OA of NL all	DK All Other	DK Gold OA	Share Gold OA of DK all	CH All Other	CH Gold OA	Share Gold OA of CH all
2000–2003	75,607	712	1%	30,616	452	1%	53,283	995	2%
2001–2004	78,087	858	1%	31,262	557	2%	54,793	1,220	2%
2002–2005	81,849	1180	1%	31,972	728	2%	56,982	1,836	3%
2003–2006	85,386	1663	2%	33,024	949	3%	60,319	2,217	4%
2004–2007	88,745	2349	3%	34,082	1,244	4%	63,205	2,790	4%
2005–2008	92,349	3265	4%	35,273	1,631	5%	65,920	3,517	5%
2006–2009	96,278	4269	4%	36,672	1,997	5%	69,518	3,912	6%
2007–2010	101,270	5587	6%	38,726	2,554	7%	72,687	4,981	7%
2008–2011	106,560	7299	7%	41,417	3,264	8%	76,658	6,354	8%
2009–2012	111,990	9504	8%	44,264	4,420	10%	80,786	7,990	10%

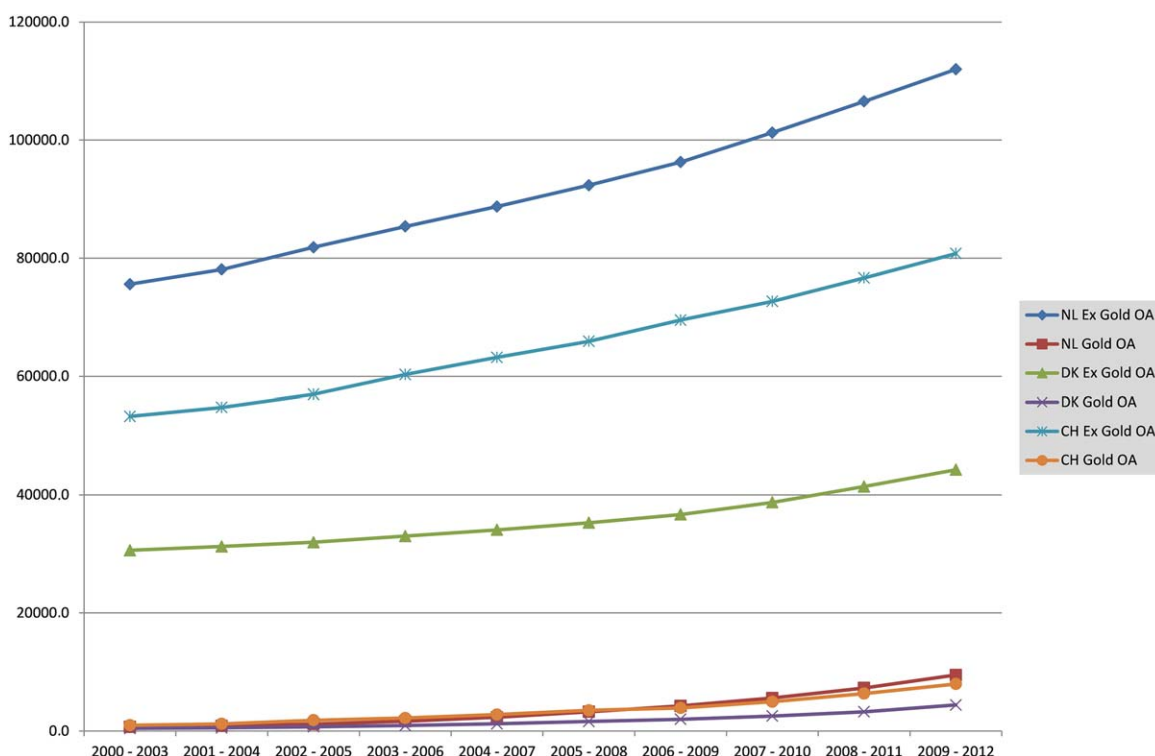
FIG. 1. Output development (P) of Denmark, the Netherlands, and Switzerland, 2000–2012/2013. [Color figure can be viewed at wileyonlinelibrary.com]

TABLE 3. Citation impact (MNCS) of Denmark, the Netherlands, and Switzerland, distinguishing Gold OA and All Other output, 2000–2012/2013.

	NL All other	NL Gold OA	DK All other	DK Gold OA	CH All other	CH Gold 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

have two ISSN numbers (journals that appear in print as well as in electronic media; for example, *Nature* appears in print under the ISSN 0028–0836, while the web version of

the journal has the ISSN 1476–4687), while it also occurs that one ISSN number is related to two journal names (this occurs in situations in which journals change name, for

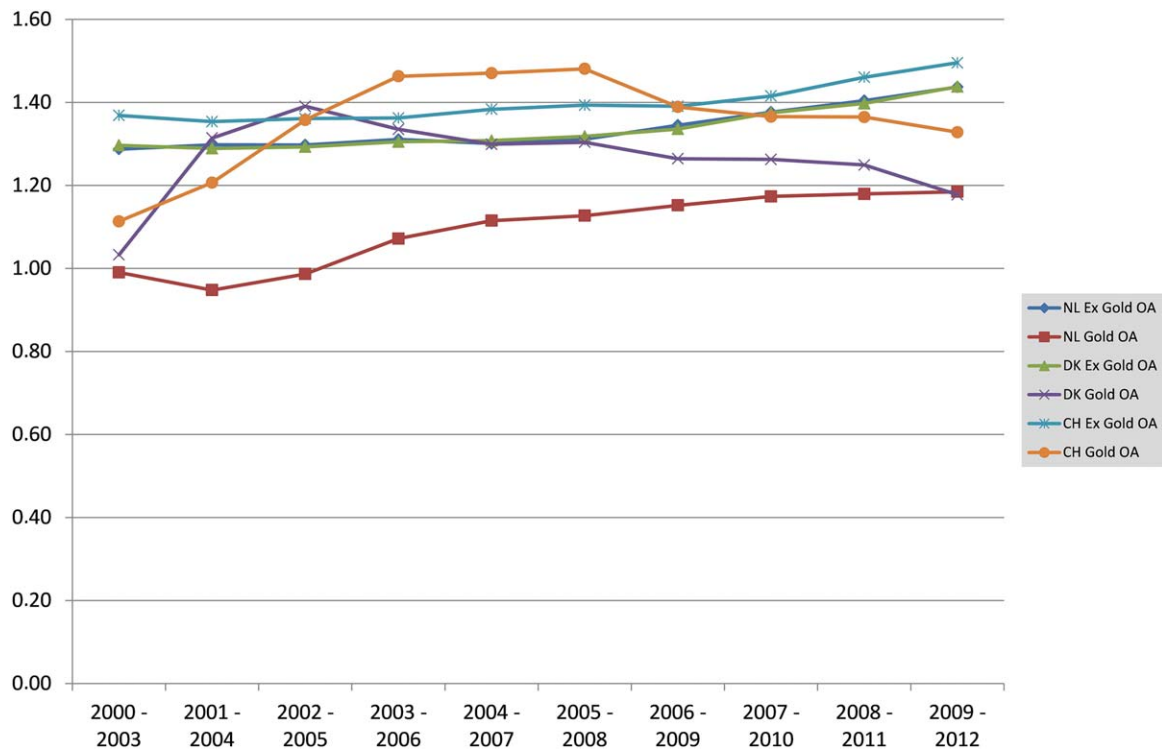


FIG. 2. Impact development (MNCS) of Denmark, the Netherlands, and Switzerland, 2000–2012/2013. [Color figure can be viewed at wileyonlinelibrary.com]

TABLE 4. Journal-to-field citation impact (MNJS) of Denmark, the Netherlands, and Switzerland, distinguishing Gold OA and All Other output, 2000–2012/2013.

	NL All other	NL Gold OA	DK All other	DK Gold OA	CH All other	CH Gold 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

example). However, that should not distort the analysis too much, as this occurs in only a very limited number of journals. So for this analysis, we used a third approach to the OA output of the three countries taking the following steps:

1. First, select within the CWTS database the output for the three countries;
2. Collect the journal's ISSN numbers, and publication years in that set;
3. Match these with the ISSN numbers and starting years downloaded from the DOAJ list;
4. Create two sets within the CWTS database, an OA formatted set of publications, and a non-OA formatted set of publications.
5. These steps were taken for all three countries, collecting publications from 2000–2013.

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 on the importance of these three types in communicating scientific findings among peers, and their relative homogeneity within the system.

Methods

In the study, we present a number of indicators: the number of publications (P); normalized article-level citation data (MNCS, Mean Normalized Citation Score), and normalized journal-level citation data (MNJS, the field normalized journal impact indicator) (Waltman et al., 2011a, 2011b). While the output indicator can be used for the various electronic systems we use in the study, and P can relate to various

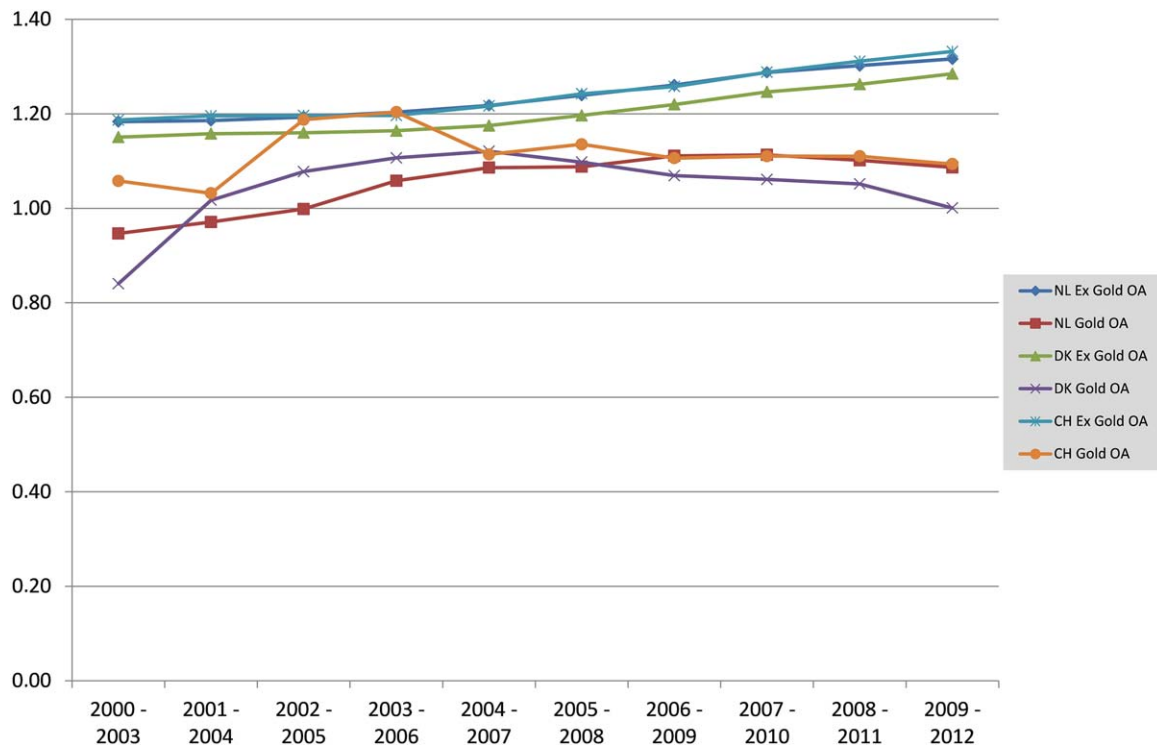


FIG. 3. Journal impact development (MNJS) of Denmark, the Netherlands, and Switzerland, 2000–2012/2013. [Color figure can be viewed at wileyonlinelibrary.com]

TABLE 5. Output (P) of Denmark, the Netherlands, and Switzerland, distinguishing Gold OA and All Other output (based on DOI-matching), 2000–2012.

	NL All other	NL Gold OA	Share Gold OA of NL all	DK All other	DK Gold OA	Share Gold OA of DK all	CH All other	CH Gold OA	Share Gold OA of CH all
2000–2003	75,607	10	0%	30,616	4	0%	53,283	2	0%
2001–2004	78,087	35	0%	31,262	25	0%	54,793	30	0%
2002–2005	81,849	136	0%	31,972	83	0%	56,982	97	0%
2003–2006	85,386	344	0%	33,024	170	1%	60,319	232	0%
2004–2007	88,745	648	1%	34,082	312	1%	63,205	420	1%
2005–2008	92,349	1,068	1%	35,273	486	1%	65,920	690	1%
2006–2009	96,278	1,531	2%	36,672	664	2%	69,518	972	1%
2007–2010	101,270	2,207	2%	38,726	924	2%	72,687	1,461	2%
2008–2011	106,560	3,036	3%	41,417	1,231	3%	76,658	2,062	3%
2009–2012	111,990	3,896	3%	44,264	1,595	4%	80,786	2,608	3%

documents types analyzed, the citation impact indicators are used only within the context of the CWTS WoS database. In the case of the impact indicators, the length of the citation window is 1 year longer than the presented year block (so, in the case of the last block, 2009–2012, the citation impact is measured up until 2013, which was the last year fully covered in the CWTS WoS database when we conducted this analysis).

Results

Results of Method I

The output numbers of the three countries according to Method I are found in Table 1. These fall into two

categories: the publications in Gold OA format and all remaining formats (which might include Green and hybrid OA publications). This is indicated by the labels “All Other and Gold OA.” The analysis covers the period 2000 until 2012 for publication data, and until 2013 for citation impact data. In this analysis we use moving publication year windows, in order to create more solid and stable trend lines. Table 1 contains the output numbers from 2000, for the three countries, and the two separate parts of the output.

The data presented in Table 2, and in particular the percentages presented therein, clearly show that Gold OA publishing is becoming increasingly important, in all three selected countries, although it remains relatively low. The

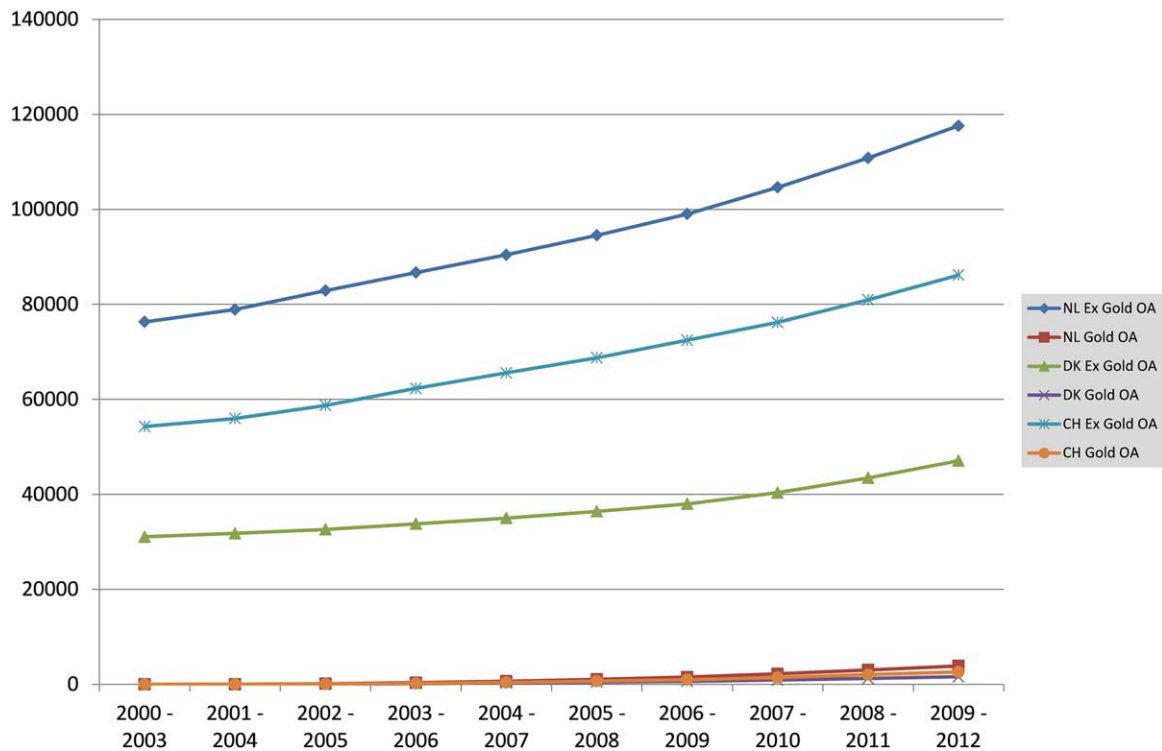


FIG. 4. Output development (P) of Denmark, the Netherlands, and Switzerland, based on matching of DOIs, 2000–2012/2013. [Color figure can be viewed at wileyonlinelibrary.com]

TABLE 6. Citation impact (MNCS) of Denmark, the Netherlands, and Switzerland, distinguishing Gold OA and All Other output (based on DOI-matching), 2000–2012/2013.

	NL All other	NL Gold OA	DK All other	DK Gold OA	CH All other	CH Gold 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

Netherlands is lagging somewhat behind Denmark and Switzerland, albeit with only a small part of the total output.

In Figure 1, we distinguish between the OA format output of the three countries (indicated by the “Gold OA” label), and the non-OA format part of the output (indicated by the “All Other” label). What we observe are increasing trends for the parts of the output not published in OA format, as well as the OA format output of these three countries. Table 2 shows that OA format output increases somewhat faster for Denmark and Switzerland as compared with the Netherlands. Clearly, OA is increasing its share of the total number of publications only very slowly and the publication system is still predominantly non-OA.

In Table 3, we present the field normalized impact (MNCS) of the outputs of the three countries, again

separated by the two types of publication output: OA and non-OA publications.

Figure 2 shows that for all three countries the “All Other” part of the output has a citation impact well above world average, with Switzerland topping the other two countries, which have a nearly equal field-normalized impact score. The impact of Gold OA publications is lower for all three countries. The impact of the Gold OA part of the national outputs of Denmark and Switzerland were initially well above the world average. This is also the case for Swiss publications, as the Gold OA format published output is lower on MNCS only from 2007–2010/2011 onwards. In the case of Denmark, this drop started somewhat earlier, while in the case of the Netherlands, the Gold OA output never got an impact higher than the “All Other” output. Another

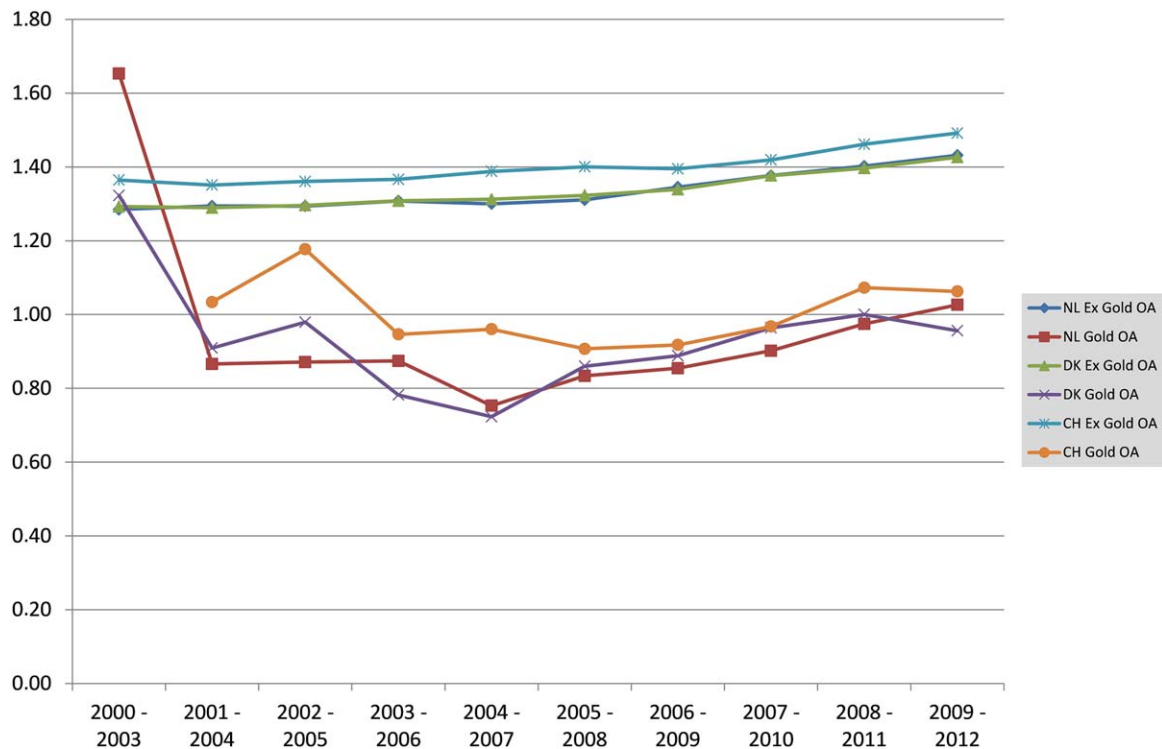


FIG. 5. Impact development (MNCS) of Denmark, the Netherlands, and Switzerland, based on matching of DOIs, 2000–2012/2013. [Color figure can be viewed at wileyonlinelibrary.com]

TABLE 7. Journal-to-field citation impact (MNJS) of Denmark, the Netherlands, and Switzerland, distinguishing Gold OA and All other output (based on DOI-matching), 2000–2012/2013.

	NL All other	NL Gold OA	DK All other	DK Gold OA	CH All other	CH Gold 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

interesting phenomenon is the increase of the gap between the impact of Gold OA and “All Other” output. This is particularly the case for Switzerland and Denmark, where we observe a clear drop of the impact of Gold OA format output compared to their “All Other” 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 4 and Figure 3), for which we use the indicator MNJS, we see an even more interesting phenomenon. While the output in “All Other” format published journals shows a choice for journals with increasing impact scores, the Gold 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: “All Other” format published journals

tend to show increasing impact scores, while Gold OA format published journals show decreasing impact trends. This is striking, since these are three of the “scientifically strong” nations, as far as can be measured with bibliometric instruments.

Results of Method II

The results of the output analysis are shown in Table 4, which again covers a similar distinction between Gold OA and “All Other” format output, but now according to the definition described earlier under Method II. We combined the DOIs of journals on the DOAJ list with the DOIs available in the WoS. From the total set of 787,611 DOIs in the DOAJ list, we matched 226,641 publications in WoS on the basis

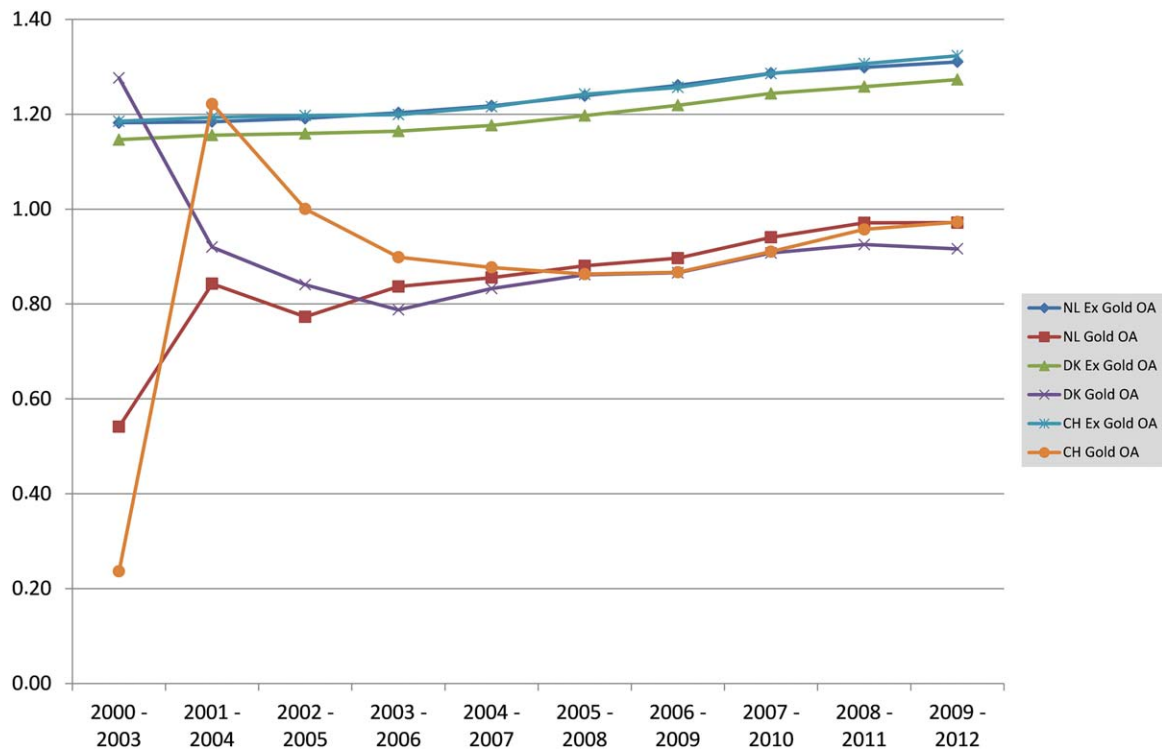


FIG. 6. Journal impact development (MNJS) of Denmark, the Netherlands, and Switzerland, based on matching of DOIs, 2000–2012/2013. [Color figure can be viewed at wileyonlinelibrary.com]

TABLE 8. Output (P) of Denmark, the Netherlands, and Switzerland, distinguishing OA and non-OA output (based on journal ISSN number and starting/publication year matching), 2000–2012.

	NL All other	NL Gold OA	Share Gold OA of NL all	DK All other	DK Gold OA	Share Gold OA of DK all	CH All other	CH Gold OA	Share Gold OA of CH all
2000–2003	75,625	681	1%	30,669	372	1%	53,340	841	2%
2001–2004	78,124	808	1%	31,312	477	2%	54,857	1,020	2%
2002–2005	81,916	1,092	1%	32,041	616	2%	57,464	1,189	2%
2003–2006	85,539	1,489	2%	33,116	797	2%	60,833	1,502	2%
2004–2007	88,928	2,140	2%	34,175	1,065	3%	63,732	1,996	3%
2005–2008	92,557	3,012	3%	35,350	1,442	4%	66,497	2,636	4%
2006–2009	96,524	3,981	4%	36,719	1,801	5%	69,745	3,313	5%
2007–2010	101,536	5,247	5%	38,760	2,327	6%	72,944	4,277	6%
2008–2011	106,877	6,848	6%	41,480	2,943	7%	77,001	5,458	7%
2009–2012	112,333	8,924	8%	44,377	3,951	9%	81,186	6,855	8%

of available DOIs. The reason for this seemingly low recall is twofold. Not all journals covered by the DOAJ list are processed for the WoS database, and not all publications in journals covered in WoS have DOIs. This means that for some journals that are both covered in the DOAJ list as well as in WoS, a match is impossible, particularly for the earlier years in the analysis. Like the first method we followed, we separated the Gold OA format published output from the Netherlands, Denmark, and Switzerland from the total set of publications for the three countries under study.

First of all, we observe that the overlap between the DOAJ list/WoS combinations with Dutch/Danish/Swiss publications in WoS is much smaller compared to the

overlap in the previous analysis of Dutch/Danish/Swiss output in OA format. This is probably the result of the missing DOIs in the WoS database. In addition, Table 5 shows lower shares of Gold OA output compared with the overall output of the three countries than Table 2. This is further underlined by Figure 4, in which the Gold OA format output of the three countries is at the low end of the graph, while we see an increase of the output of the “All Other” format output of the three countries.

In Table 6, we present the impact scores of the three countries, again distinguishing Gold OA format output and “All Other” format output. Again, we observe lower impact scores for the Gold OA format output of the three countries, except for the starting block of the analysis (please note that

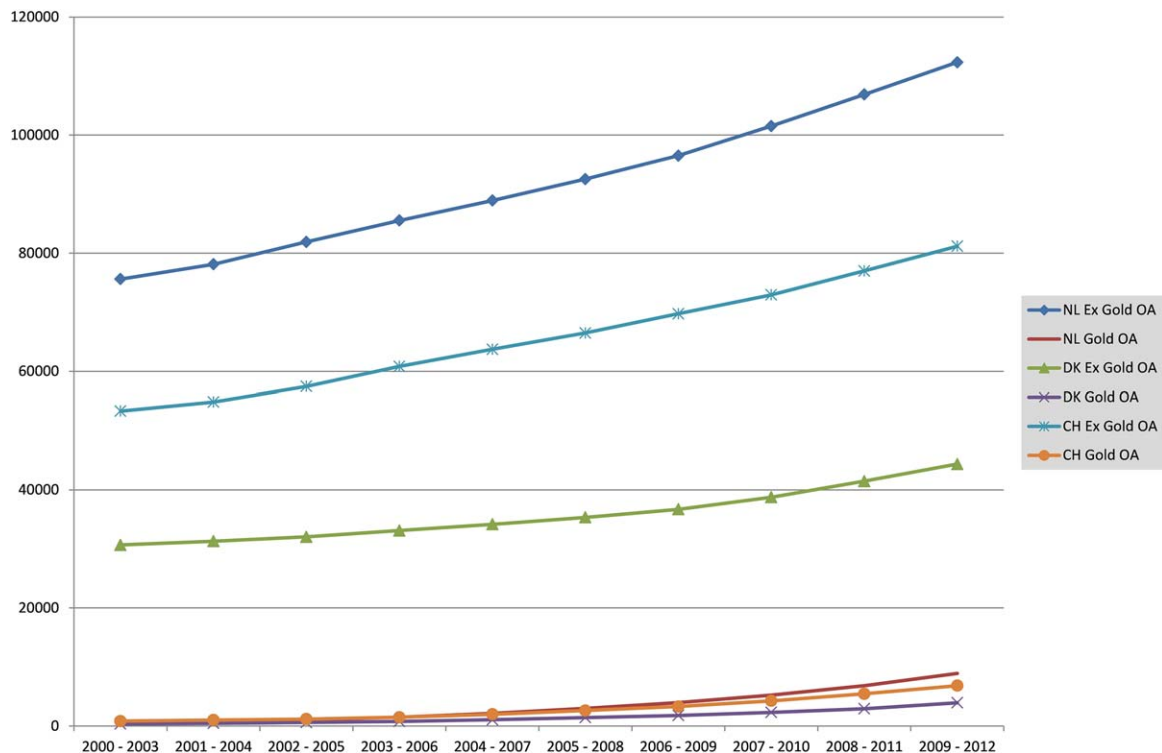


FIG. 7. Output development (P) of Denmark, the Netherlands, and Switzerland, based on journal ISSN number and starting/publication year matching, 2000–2012/2013. [Color figure can be viewed at [wileyonlinelibrary.com](#)]

TABLE 9. Citation impact (MNCS) of Denmark, the Netherlands, and Switzerland, distinguishing Gold OA and “All Other” output (based on journal ISSN number and starting/publication year matching), 2000–2012/2013.

	NL All other	NL Gold OA	DK All other	DK Gold OA	CH All other	CH Gold OA
2000–2003	1.29	1.02	1.30	1.06	1.36	1.34
2001–2004	1.30	0.94	1.29	1.39	1.35	1.30
2002–2005	1.30	0.97	1.29	1.48	1.36	1.40
2003–2006	1.31	1.03	1.30	1.42	1.36	1.38
2004–2007	1.30	1.10	1.31	1.32	1.39	1.30
2005–2008	1.31	1.11	1.32	1.28	1.40	1.27
2006–2009	1.35	1.14	1.34	1.21	1.39	1.27
2007–2010	1.38	1.15	1.38	1.20	1.42	1.27
2008–2011	1.41	1.16	1.40	1.19	1.47	1.27
2009–2012	1.44	1.18	1.44	1.11	1.50	1.27

the output numbers are extremely low in this part of the analysis for the Netherlands and Denmark: respectively 10 and 4 articles). From the second year block onwards, we observe increasing trends in the impact of the Gold OA format of the three countries, although we must stress that this is also the case for the “All Other” format output of the three countries.

Figure 5 shows this development of the impact scores of both sets of publications. The impact scores of both sets are increasing, although the difference between OA and non-OA remains more or less the same.

In Table 7 we present the outcomes of the analysis on the journal impact scores, based on Method II. Here we observe, similar to the previous outcomes, fluctuations in the initials years of the analysis for the Gold OA format output,

followed by a more stable situation from 2005–2008 onwards. This is even more visible in Figure 6.

Results of Method III

The results of the output analysis are shown in Table 8, which covers a similar distinction between Gold OA and “All Other” format output, but now according to the definition described earlier under Method III. So we matched the data sets from WoS with the publications in the journals on the DOAJ list on the basis of the ISSN numbers, and in addition to that, assuming that starting year on the DOAJ list is similar to the publication year in WoS, a “year” similarity.

The numbers of publications resulting from this method are clearly higher as compared with the previous method, in

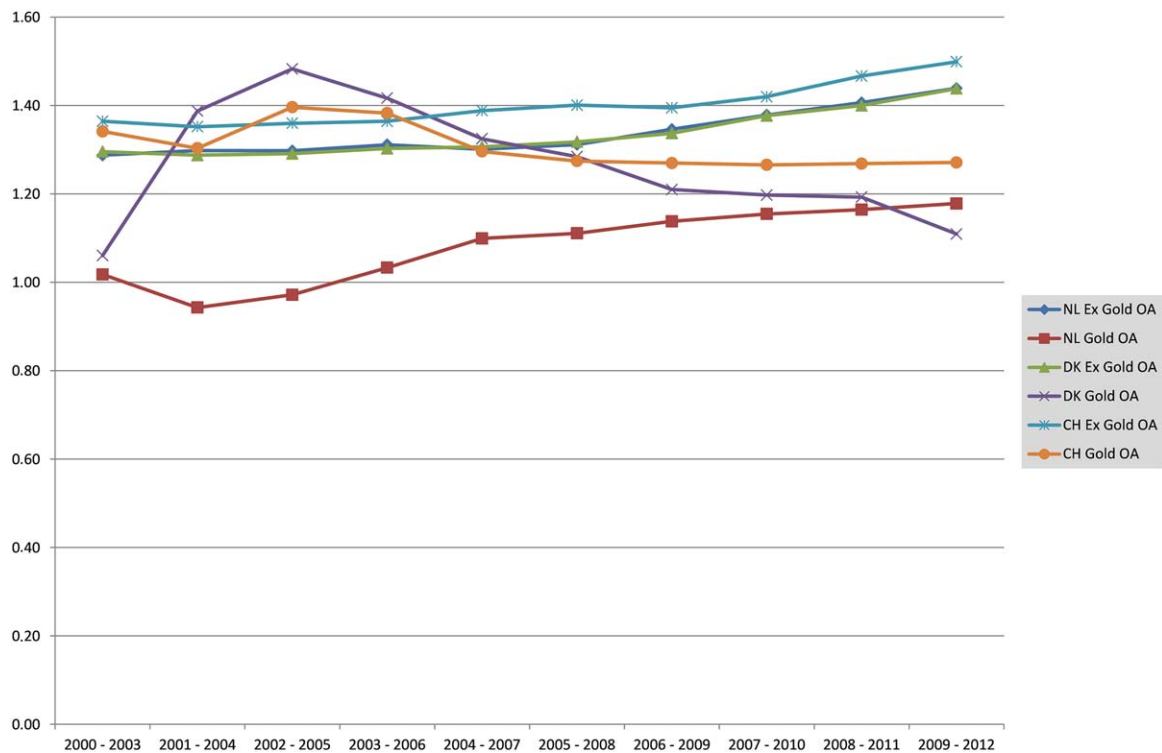


FIG. 8. Impact development (MNCS) of Denmark, the Netherlands, and Switzerland, based on journal ISSN number and starting/publication year matching, 2000–2012/2013. [Color figure can be viewed at wileyonlinelibrary.com]

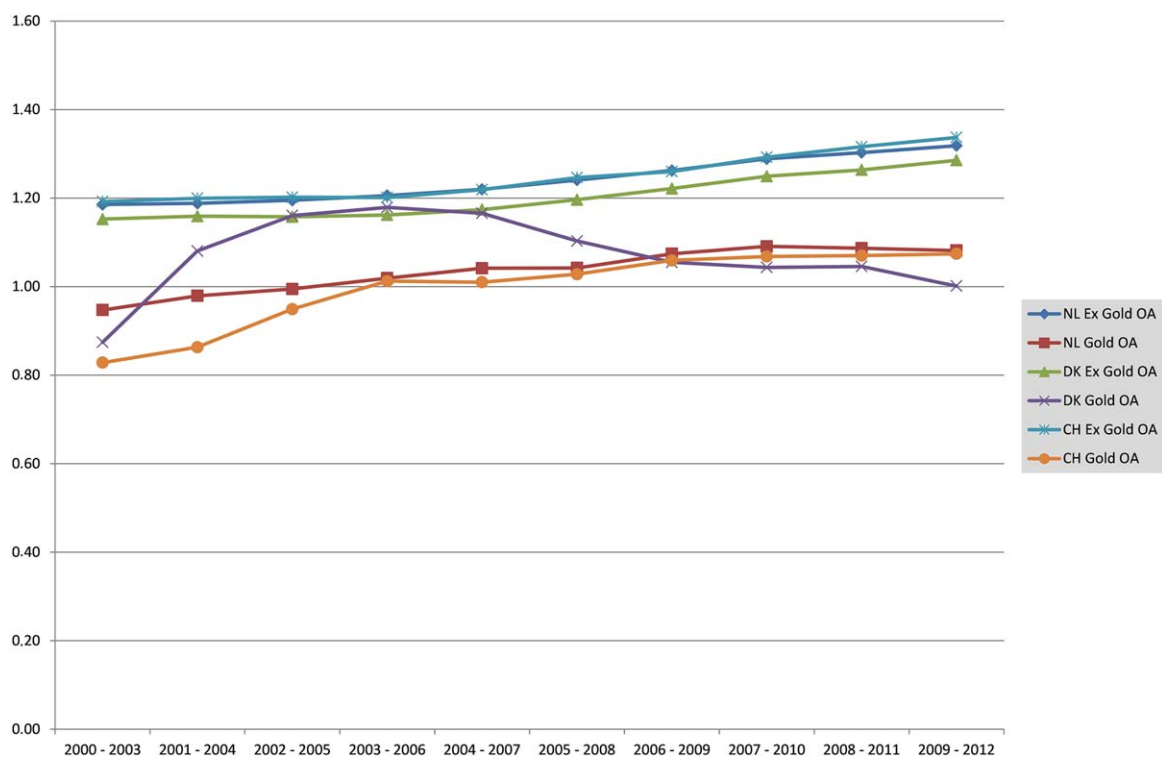


FIG. 9. Impact development (MNJS) of Denmark, the Netherlands, and Switzerland, based on journal ISSN number and starting/publication year matching, 2000–2012/2013. [Color figure can be viewed at wileyonlinelibrary.com]

TABLE 10. Citation impact (MNJS) of Denmark, the Netherlands, and Switzerland, distinguishing Gold OA and “All Other” output (based on journal ISSN number and starting/publication year matching), 2000–2012/2013.

	NL All other	NL Gold OA	DK All other	DK Gold OA	CH All other	CH Gold OA
2000–2003	1.19	0.95	1.15	0.87	1.19	0.83
2001–2004	1.19	0.98	1.16	1.08	1.20	0.86
2002–2005	1.19	0.99	1.16	1.16	1.20	0.95
2003–2006	1.21	1.02	1.16	1.18	1.20	1.01
2004–2007	1.22	1.04	1.17	1.17	1.22	1.01
2005–2008	1.24	1.04	1.20	1.10	1.25	1.03
2006–2009	1.26	1.07	1.22	1.05	1.26	1.06
2007–2010	1.29	1.09	1.25	1.04	1.29	1.07
2008–2011	1.30	1.09	1.26	1.05	1.32	1.07
2009–2012	1.32	1.08	1.29	1.00	1.34	1.07

which the DOAJ list was used as well. We solved the issue of the missing DOIs by matching by ISSN and year. The numbers are somewhat similar to the results derived from Method I (as one might expect, given the definition of WoS OA disclosure, in which DOAJ also plays a role), although they are somewhat lower in Method III compared to Method I. This is probably due to the fact that the method of data collection underlying Method I also included hybrid OA publications: OA publications in journals that are otherwise not OA (yet).

The results of Methods I and III presented in Tables 2 and 8 are similar, and Figures 1 and 7 show similar trends.

Table 9 contains the MNCS scores of the three countries, of both the Gold OA and the “All Other” format output. The normalized impact scores (MNCS) for the “All Other” format output shown in Table 9 and Figure 8 are nearly exactly the same as the ones presented in Table 2 and Figure 2. The differences observed in the MNCS scores of the Gold OA format output can be explained by the presence of hybrid publications in the Gold OA format output underlying Method I (and consequently, in the tables and figures related to this method). Apparently, OA publications in hybrid journals have a higher impact than the non-OA output of those journals, as the journals that follow the hybrid business model often exist for many years, and have very strong reputations, contrary to many new OA journals.

In Figure 9, the decrease of the MNCS values for the Gold OA format output for the three countries starts earlier than in Figure 2. Finally, MNCS scores are lower than the scores based on Method I.

In Table 10 we present the MNJS scores of the journals related to the Method III approach. Compared to the results in Table 3, we notice an even stronger divergence of MNJS values of the Gold OA and the “All Other” format output of the three countries in the study.

Conclusion and Discussion

We now summarize the main findings and discuss the limitations in the ways OA is disclosed in electronic systems supporting bibliometric analyses. Finally, we discuss the need to improve identification of OA publications and the use of bibliometric techniques to measure OA.

Our conclusions are limited to the domains in which journal publishing is the dominant way of communication (the natural, life, and medical sciences, and to a lesser extent the social sciences and humanities; van Leeuwen, 2013). We observe for the three countries that the share in output in Gold OA journals is smaller as compared to the remaining output per country. We observe a divergence in the development of citation impact for (Gold) OA and all other publications with consistently lower impact for the Gold OA publications.

Gold OA journals have lower journal impact scores than all other journals. This may mean that they still struggle to find their position within the total “reputational hierarchy” of the domain. This is a common problem for new journals and (Gold) OA journals are no exception to that rule. Our findings are consistent with the results of other studies: Gold OA is not associated with a citation advantage, nor with a disadvantage (for example, Archambault et al., 2014). With the inclusion of the various forms of Green OA, we would expect to find a larger proportion of OA articles and a more nuanced outcome related to impact. That Green OA has been found to have increased accumulation of citations (Archambault et al., 2014) may be associated with the circumstances identified earlier as confounding factors (for example, early exposure, multiple access points, and proximity of researchers).

Our results indicate that we may need to worry about the role of peer review in the journals that are part of the expansion of the WoS database in the last couple of years, many of which are in the OA segment of the database. The Institute for Scientific Information, the predecessor of the current owner of the WoS database, Clarivate Analytics always 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, regular appearance, preferably in the English language, etc.). We do not know whether this is still such a strong criterion, particularly given the fact that so many new journals appeared around the OA development.

Our study also shows that the various manners by which OA is defined in electronic databases do not follow comparable criteria. The ways two main formats of OA can be operationalized within the world of WoS is an example of this unclear and somewhat messy situation. The fact that the

Scopus database did not have the functionality to clearly define OA for users of the system is another instance of the situation around OA.⁴ A further expression of this lack of clarity is the various ways OA 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 OA). Yet another example relates to the various license types related to OA.

The increased use of CRIS systems used at research institutions provides an alternative approach to monitoring research output. Rather than relying primarily on commercial data sets, such as WoS and Scopus, CRIS systems can be used to build more complete data sets through local data collection. This enables coverage of output types not included in commercial data sets (Sivertsen, 2014). A recently published metadata standard for OA (Carpenter, 2013) holds some promise for improving identification of OA in conjunction with the use of CRIS systems. Here too, however, stakeholders involved in the new standard were unable to agree on a precise definition of OA. Instead, the standard specifies metadata elements for *free to read* and *license reference*, the latter of which should point to copy-right information publicly accessible on the web (NISO, 2015). More standardization of metadata is needed to reliably identify all types of OA publishing. In the meantime, increased attention to national research assessment and increased use of institutional CRIS systems together provide a potentially welcoming context for implementing new metadata practices. This would ideally include the possibility of tracking OA among the diversity of research outputs maintained by CRIS systems and considered in assessment events. In this context, it becomes important to assign openly accessible, persistent identifiers to all research objects (Tatum & Wouters, 2014). This could increase the potential of institutional research information for tracking OA as part of regular research assessment practices, rather than relying on the present approach of estimation derived from random sampling of commercial data sets.

References

- Archambault, E., et al. 2014. "Proportion of Open Access Papers Published in Peer-Reviewed Journals at the European and World Levels—1996–2013." *Rapport, Commission Européenne DG Recherche & Innovation; RTD-B6-PP-2011-2: Study to Develop a Set of Indicators to Measure Open Access*.
- Björk, B.-C. 2012. The hybrid model for open access publication of scholarly articles: A failed experiment? *Journal of the American Society for Information Science and Technology* 63(8): 1496–1504.
- Björk, B.-C., et al. 2010. Open access to the scientific journal literature: Situation 2009. *PLoS One* 5(6): e11273.
- Björk, B.-C., Laakso, M., Welling, P., & Paetau, P. 2013. Anatomy of green open access. *Journal of the Association for Information Science and Technology* 65(2): 237–250.
- BOAI. 2002. Budapest Open Access Initiative. *The Open Society Foundations*. <http://www.opensocietyfoundations.org/openaccess>.
- Carpenter, T. 2013. Progress toward open access metadata. *Serials Review* 39(1): 1–2.
- Davis, P.M., & M.J. L. Connolly. 2007. Institutional repositories: Evaluating the reasons for non-use of Cornell University's Installation of DSpace, March. <http://hdl.handle.net/1813/5195>.
- Kurtz, M.J., et al. 2005. The effect of use and access on citations. *Information Processing & Management, Special Issue on Infometrics*, 41(6): 1395–1402.
- Laakso, M., & Björk, B.-C. 2013. Delayed open access: An overlooked high-impact category of openly available scientific literature. *Journal of the American Society for Information Science and Technology* 64(7): 1323–1329.
- Lee, K., Brownstein, J.S., Mills, R.G., & Kohane I.S. 2010. Does collocation inform the impact of collaboration? *PLoS One* 5(12): e14279.
- Moed, H.F. 2007 The effect of "Open access" on citation impact: An analysis of ArXiv's condensed matter section, *Journal of the American Society of Information Science & Technology* 58(13), 2047–2054.
- NISO. 2015. *Access License and Indicators NISO RP-22-2015*. National Information Standards Organization. ISBN: 978-1-937522-49-0
- Sivertsen, G. 2014. Scholarly publication patterns in the social sciences and humanities and their coverage in Scopus and Web of Science. In *Proceedings of the Science and Technology Indicators Conference 2014*. Leiden, the Netherlands.
- Swan, A. 2010. The open access citation advantage: Studies and results to date. Technical Report. <http://eprints.ecs.soton.ac.uk/18516/>.
- Tatum, C., & Wouters, P.F. 2014. Next generation research evaluation: The ACUMEN Portfolio and Web Based Information Tools. In *OpenAIRE-COAR Conference: Open Access Movement to Reality—Putting the Pieces Together*. Athens, Greece.
- van Leeuwen, T.N. 2013. Bibliometric research evaluations, Web of Science and the Social Sciences and Humanities: A problematic relationship? *Bibliometrie Praxis und Forschung* 1–18 (<http://www.bibliometrie-pf.de/article/viewFile/173/215>)
- Waltman, L., van Eck, N.J., van Leeuwen, T.N., Visser, M.S., & van Raan, A.F.J. 2011a. Towards a new crown indicator: Some theoretical considerations, *Journal of Informetrics* 5(1): 37–47
- Waltman, L., van Eck, N.J., van Leeuwen, T.N., Visser, M.S., & van Raan, A.F.J. 2011b. Towards a new crown indicator: An empirical analysis. *Scientometrics* 87(3): 467–481
- Xia, J., Myers, R.L., & Wilhoite, S.K. 2011. Multiple open access availability and citation impact. *Journal of Information Science* 37(1): 19–28.

⁴In the design of the study commissioned by the Dutch Ministry of Science, Culture & Education (OC&W), Scopus was also considered as a data source for conducting this analysis.