The deliverance of open access books

Examining usage and dissemination

Ronald Snijder

The deliverance of open access books

For Dorien and Charlotte

The deliverance of open access books

Examining usage and dissemination

Ronald Snijder

ISBN 978-90-8555-120-1
NUR 615

‘Gladys, the thing about books... well, the thing... I mean, just because it’s written down, you don’t have to... that is to say, it doesn’t mean it’s... what I’m getting at is that every book is – ‘
He stopped. They believe in words. Words give them life. I can’t tell her that we just throw them around like jugglers, we change their meaning to suit ourselves...
He patted Gladys on the shoulder. ‘Well, read them all and make up your own mind, eh?’
Making Money / Terry Pratchett, 2007

Too much information, and so much of it lost. An unindexed Internet site is in the same limbo as a misshelved library book. This is why the successful and powerful business enterprises of the information economy are built on filtering and searching.
The Information : a History, a Theory, a Flood / James Gleick, 2011

Many, many thanks

This publication could only exist through the generous support of many people, and I would like to express my heartfelt thanks.

Eelco Ferwerda triggered all this, by asking me – in 2008 – to look into the role of open access on books. While I still feel I have not completely answered his question, at least we know more since then.

Professor Paul Wouters has guided and challenged me throughout the PhD project, which started in 2011. Many of the improvements stem from his patience and knowledge. Since 2015, Professor Frank Huysmans has been an equally supportive mentor. Most of the chapters have been published – in open access – and this has been made possible by several publishers, copy-editors and peer reviewers.

Much of my research revolved around the OAPEN Library and the Directory of Open Access Books. Many people were directly involved: Lotte Kruijt, Caspar Treijtel, Hans Gommers, Hans Scholte, Salam Baker Shanawa, Janneke Adema and Paul Needham among many others.

Assistant Professor Lucy Montgomery and Alkim Ozaygen have been more than generous with their time.

My colleagues at Data Office UWV and my former colleagues at Amsterdam University Press have always been forthcoming, enabling me to juggle two jobs and this project. I am also grateful for the assistance from CWTS. Rob Wadman helped to make this document presentable.

My old friend Diebert van Rhijn acted as my IT advisor. All my friends and my family helped me by not minding my obsession with this seemingly endless project.

But most of all I want to thank the two most important people in my life: my wife Dorien and my daughter Charlotte. Without them, I would not have been able to carry out this crazy project. For this reason – and many, many more – I dedicate this publication to them.

1 Contents

2 Introduction

This publication will discuss the dissemination and usage of open access monographs, something that I have been working on since 2008. Here, open access monographs are defined as a scholarly piece of writing of book length on a specific subject, disseminated online in such a way that its contents can be read and downloaded without any barrier. Disseminating academic books in this manner is part of the open access movement, which aims to make scientific and scholarly content available to all. Peter Suber – considered to be the de facto leader of the open access movement – describes the rationale as such: "[R]esearch that is worth funding or facilitating is worth sharing with everyone who can make use of it." (Suber, 2012).

Platforms for open access monographs are fairly new and they are just one aspect of the changes in the way scholarly and scientific results are made public. As I became involved in the development of both the OAPEN Library and the Directory of Open Access Books, questions on optimization arose. How can we improve these open access book platforms if there are few examples to learn from? An optimal solution should be based on evidence and my research on the dissemination and usage of freely available academic books aims to uncover relevant facts.

2.1 A short history of open access

Starting in 1991, preprints of physics papers were distributed using a central repository mailbox. The number of articles grew, and the repository expanded to include astronomy, mathematics, computer science, quantitative biology. In 2001, this repository was renamed to arXiv.org. The rise of the world wide web further enabled worldwide online distribution and in 2002 this idea was captured in the Budapest Open Access Initiative (BOAI) declaration (Chan *et al.*, 2002), where the term "Open Access" was coined. In the same year, the first set of Creative Commons licenses was released. These licenses enable the reuse of the contents in varying degrees. The role of licenses in the dissemination of open access books will be discussed in more detail in chapter 5.

At the start of the twenty-first century, several large scale open access initiatives were founded: PubMed Central[1] and the Public Library of Science.[2] Since then, other journal platforms such as PeerJ,[3] F1000Research[4] and Open Library of Humanities[5] have emerged. An important online book platform, the Google Books program, started in 2002.[6] A decade later saw the launch of several open access monographs platforms. In 2010, the OAPEN Library[7] was launched. In 2012, the Directory of Open Access Books[8] was introduced, listing monographs contained on several platforms. The next year, SciELO[9] and OpenEdition[10] started book platforms.

The introduction of new platforms for journal articles and books is part of a profound change in scholarly communication: the traditional roles of participants are changing. Some publishers are building their own digital collections, a task normally associated with libraries. On the other hand, academic libraries are starting up publishing activities (Bonn & Furlough, 2015), and publishers like Open Book Publishers or the Open Library of Humanities are led by academic authors. Lastly, some funders are managing their own collections, and – through crowdfunding – readers can finance books. For instance, the Austrian science fund FWF directly places books in the OAPEN Library (Snijder, 2015). Other funding bodies – such as the Spanish National Research Council – have chosen to set up an institutional repository (Bernal, 2013). The organisation Unglue.it uses a crowdfunding model to pay the rights holders of books to make them available through an open license. Among other types of books, academic books are part of the crowdfunding efforts (Howard, 2012).

2.2 Defining usage

Providing a general definition of "usage" is challenging; in this publication, the term "usage" as it refers to open access monographs is defined

1 https://www.ncbi.nlm.nih.gov/pmc/
2 https://www.plos.org/
3 https://peerj.com/
4 https://f1000research.com/
5 https://www.openlibhums.org/
6 https://www.google.com/intl/en/googlebooks/about/history.html
7 https://www.oapen.org
8 https://www.doabooks.org
9 http://books.scielo.org/
10 http://books.openedition.org/

as accessing the contents of the books. This is not exactly the same as reading a monograph. Most of this publication's research is done using the OAPEN platform. On that platform, it is not possible to measure whether a monograph has been read; instead the number of downloads is recorded. In a similar vein, the usage of the Google Book platform is measured as the number of pages that have been shown, or the number of times a book has been accessed. The results of the OAPEN Library and Google Books can be seen as a proxy for reading the books, but "flipping" a page in Google Books or downloading a book from the OAPEN library is no absolute guarantee that the person has actually read the monograph.

Many open access advocates stress the importance of reusing the contents of the scientific or scholarly documents that have been made available freely. This is supported by open licenses such as the Creative Commons licenses, which enable a certain amount of reuse by others. While the importance of reuse is not disputed, I will not discuss it in much detail. The primary reason is that reuse is even harder to measure than accessing content. At this very early stage in the development of open access monograph platforms, there are no reliable indicators available. This is not limited to reuse. For journal articles, measuring the number of citations is common practice. For monographs, this is not the case: chapter 9 describes the difficulties to obtain citations. Thus, in my definition of usage I have purposefully omitted reuse.

The question whether open access leads to more usage of monographs has already been settled in other research (Emery *et al.*, 2017; Ferwerda, Snijder, & Adema, 2013; Snijder, 2010). Making academic books freely accessible invariably increases the number of pages read online or the number of copies downloaded; a conclusion that is rather obvious. The next phase is to examine how to optimize that usage, and whether the increased usage has positive effects in academia and beyond.

The dissemination of open access monographs depends on platforms that offer a two-parts solution: a digital collection and the means of dissemination. When a platform is created, its administrators have to make decisions on what books to include. The collection as a whole will affect which users will be interested in using the platform, but we will also see that different aspects of the individual books affect the usage. Throughout the publication, the role of subject and language will be discussed in detail.

However, whether the platform reaches the intended audience depends not only on its contents. Just as important are the technical possibilities of the platform. Not just the question of how visitors can interact with the platform is significant, but also whether the contents can be integrated

into other environments. The impact of content integration will be made visible in chapter 4 and chapter 5.

2.3 Books versus journals

The difference in coverage between articles and monographs is visible in a recent review article on the impact of open access (Tennant *et al.*, 2016). It aims to list all current knowledge of this subject, but focuses only on journal articles as a way to publish scientific or scholarly results. However, monographs are an important publication type in the humanities and social sciences. Williams *et al.* (2009) conclude that "the monograph continues to enjoy unique appeal and status", a clear indication of its standing.

Journal articles and monographs differ in several ways. The most obvious difference is the length: the average number of pages in an article is most likely around fifteen,[11] while the average monograph will contain around 300 pages. The latter publication form is clearly more suited for a thorough discussion of a subject. However, a longer text also changes the preferred format: while articles are mostly read digitally, there is still demand for paper books. In this light, it is understandable that publishers and librarians are interested in the combination of open access and paper versions. Chapter 3 describes my research into the influence of open access on the sales of paper copies.

The number of book titles and the number of journal articles differ wildly. This is illustrated by the Directory of Open Access Journals (DOAJ) and the Directory of Open Access Books (DOAB). In August 2017, the DOAJ lists over 2.5 million articles. In contrast, the DOAB contains close to 8,900 titles. This difference has economic consequences. Articles tend to be more standardized, and due to concentration of publishers, economies of scale can be more easily achieved. In contrast, monographs tend to be treated like unique projects, and are published by a much larger number of publishers, considerably differing in size.

The difference in text length also leads to a different pace of interaction: it takes longer to write a monograph than it takes to create an article. Using citation analysis based on what is common in journal articles will not lead to optimal results. Any citation analysis on academic books, such as the research in chapter 9, has to accommodate for this. If the long "citation cycles" are problematic, other forms of assessment might be examined: for

11 See for instance Falagas, *et al.* (2013); Stremersch, Verniers, & Verhoef (2007)

instance, by looking at the usage data of open access monographs. This idea is further investigated in chapter 7 .

2.4 Central thesis and research questions

In the introduction, I described my involvement with the OAPEN Library and the Directory of Open Access Books. Ultimately, these platforms aim to share the contents of freely accessible books as widely as possible, which is measured by the level of usage. While the usage of open access monographs depends on the removal of paywalls, the level of usage is primarily determined by other factors. Properties of the books such as language and scholarly field determine the possible readers and the way dissemination platforms are configured affect whether those readers can actually be reached.

The question which factors affect the use of open academic books is quite open-ended. In this publication, I will examine three main aspects: economic sustainability, optimisation of the infrastructure and evaluation of the results. Economic sustainability of open access monograph publishing is one of the basic conditions for the platforms: without books, there is no need for a platform. This leads to the question whether open access has a positive influence on the sale of monographs. For decades, the uneasy financial situation surrounding publishing academic books has been known as "the monograph crisis". Decreasing sales and rising costs are threatening the economic sustainability of monograph publishing and publishers are exploring alternative business models. One of these is the so-called "hybrid model", where an online version is made freely available, and paper copies must be purchased. Will the improved visibility lead to more sales? This is explored in chapter 3.

In addition to the economic aspects, I have examined the factors affecting the dissemination of open access monographs. Understanding these factors helps to optimize the platforms. A fundamental question for the development of both the OAPEN Library and DOAB is how to present the collection to prospective readers. Should the platform only be accessible as a "silo", or should it try to integrate its offering in other systems? The answer to this question has consequences for the design. The "silo" approach assumes that humans reach the platform and start searching there, while system integration requires standardized book metadata that can be imported into the systems of libraries and aggregators. Chapter 4 deals with this question.

The optimization of open access platforms is not just dependent on technical choices. Another thing to consider is collection choices. In the case of the OAPEN Library, the collection contains books with a license that permits reuse and books with a more restrictive license. Does this difference in licensing affect the use? I have compared the usage of the two sets of books. Within the open access community, licenses that enable reuse of scholarly content are seen as very important, and chapter 5 examines whether this preference is also shared by the users of the OAPEN Library. Furthermore, the influence of aggregation through another platform – the Directory of Open Access Books – is measured.

Apart from licenses, the users of the OAPEN Library and DOAB may have other preferences. Understanding those preferences is useful to improve the platforms, but users are not required to register. Thus, no information about individuals is stored. The question is how to emulate the successful tactics of online retailers – that store the preferences of their clients – without violating privacy. A possible solution can be found in deploying social analysis techniques to discover user communities. See chapter 6.

The next chapters discuss the results of open access monographs dissemination; starting with the question of how to evaluate the effects of open access monographs. I have examined the possibility to quantify the effects of Humanities and Social Sciences (HSS) research, in a way that is relatively effortless. The results of chapter 7 aim to achieve this goal through the investigation of usage data.

If the goal of open access is to make research available to everybody, does it help to overcome the digital divide between the "global north" and the "global south"? Using open access book platforms requires a functioning digital infrastructure, which might set back readers in developing countries. Does open access lead to more usage in developing countries? The answer can be found in chapter 8. Lastly, I have examined if there is an "open access advantage" for monographs. It is widely documented for journal articles, but the effect of open access on citations is largely unknown. The same holds true for social media. Chapter 9 discusses the question whether open access monographs are cited more and receive more attention on social media.

I have conducted multiple studies on the usage of open access monographs, which are presented in the following chapters. Each chapter reviews a different aspect: book sales, digital dissemination, open licenses, user communities, measuring usage, developing countries and the effects on citations and social media.

3 The influence of open access on monograph sales : The experience at Amsterdam University Press

Snijder, R. (2014). The Influence of Open Access on Monograph Sales : The experience at Amsterdam University Press. LOGOS: The Journal of the World Book Community, 25(3), 13–23. https://doi.org/10.1163/1878-4712-11112047

3.1 Introduction

For years, decreasing sales have threatened the sustainability of monograph publishing and this has led to a search for alternative models. Most of these proposed models are hybrid: they contain an open access component combined with selling other versions of the book. In this paper, the experiences of Amsterdam University Press (AUP) in using a hybrid model are analysed by looking at the effect of open access publishing on monographs sales. However, several other influences may also affect sales, and these will be taken into account as well. The goal is to find what effect making books freely available online has on sales. To achieve this, I shall apply statistical methods to the sales data of 513 books from a period of three years.

The economic problems concerning monographs have been discussed by Wasserman and Thompson. Wasserman (1998) discusses the costs of publishing monographs and the dramatic effects of declining library sales. Thompson (2005) extensively reviews the challenges – including financial challenges – facing monograph publishing.

Others look at the possibilities of digital publishing in an open access model. Greco & Wharton (2008) conclude that university presses cannot survive on a 'print-only' business model and should consider open access publishing. Steele (2008) draws more or less the same conclusion and describes the open access model as 'a viable alternative when placed within institutional settings'. Houghton *et al.* (2009) discuss the costs of scholarly publishing – including the costs of monographs – and conclude that open access publishing is beneficial for society. Withey *et al.* (2011) acknowledge a trend towards more open access publishing, but stress the need for sustainable business models. Cross urges academic libraries to support small

academic publishers by purchasing their monographs (Cross, 2011). Pinter (2012) also discusses the financial perils of publishing monographs and proposes a solution in which a consortium of libraries fund an open access version of a title, enabling the publisher to sell enhanced digital or paper versions of the book.

Recently, Ferwerda (2014) listed the current business models for open access and monographs, ranging from a hybrid publication model to crowdfunding. Jackson (2014) – a publisher at Oxford University Press – describes the current lack of demand for a publishing model in which all costs are met before publication. At this moment, there is no consensus regarding a 'proven' business model.

Some authors also try to find evidence of whether free digital versions of a book have an effect on sales. Hilton & Wiley (2011) conclude that a correlation exists between a free e-book and increased print sales. Their research used an experimental group of eight books and a control group of six books, both fiction and non-fiction. Snijder (2010) set up an experiment on monographs, using three experimental groups of 100 title each and a control group of 100 titles. One result was that making a book freely available did not affect the number of copies sold. Based on the same principles, the Dutch-based OAPEN Foundation set up a two-year experiment: OAPEN-NL. During that period, 50 books were published on open access and also as a paper monograph. Several aspects – sales among them – were monitored and compared with a control group of comparable titles published in the traditional way. The results were similar to Snijder's results: the number of copies sold was not affected by publishing on open access (Ferwerda *et al.*, 2013). In the UK, JISC set up an experiment called OAPEN-UK. Here, 29 'matched pairs' of monographs are compared: one title in each pair is made available on open access while no changes are made to the other's publication model (Collins & Milloy, 2012).

This paper does not follow the same controlled arrangement used by Snijder (2010) or the OAPEN-NL experiment. Instead of investigating carefully balanced data sets, I use all titles published under one imprint published by AUP. Whereas the experiment of 2010 used data selected over nine months of the year 2009, here the time frame is much larger: 36 months, the years 2010 to 2012.

Amsterdam University Press is an academic publisher – owned by the University of Amsterdam – that publishes monographs and journals, mostly in the field of humanities and social sciences (AUP, 2012). The Press has gained extensive experience with open access publishing. The open access monographs are always made available via a hybrid model in which the

print version of the book is sold and a digital version is made available free. Since 2010, the open access titles published under the imprint 'Amsterdam University Press' have been released not only through AUP's repository but also via the OAPEN Library. The OAPEN Library is an important dissemination channel for AUP's open access books: in April 2014, the Library contained 447 titles published by AUP.

The OAPEN Library (http://www.oapen.org) was officially launched in September 2010 (OAPEN Consortium, 2011). It is a web-based collection of open access monographs published by dozens of publishers. In April 2014, the collection contained over 2100 titles by 68 publishers. The OAPEN Library offers several ways to access its contents: it enables searching and browsing, readers can share book descriptions via social media, and it contains several data feeds (Snijder, 2013a). Amsterdam University Press is part of the board of the OAPEN Foundation, which maintains the OAPEN Library.

3.2 The data set

In this paper, the following research question will be discussed: what is the influence of open access on monograph sales and how large is the influence of open access publishing compared with other influences on monograph sales? The data set consists of 513 books published under the imprint 'Amsterdam University Press'. All books published under this imprint are subject to peer review. The group of books consists of 69 published in 2010, 68 published in 2011, 62 published in 2012, and 319 published between 1995 and 2009. Over 70 per cent of those books – 378 titles – were published on open access and are available in the OAPEN Library (Table 1).

Table 1. Titles available and not available on open access

	Number of titles	*Percentage*	*Number of copies sold*	*Sales percentage*	*Average sales per title*
On open access	378	73.7	67 210	65.6	66.3
Not on open access	135	26.3	35 170	34.4	105.6
Total	513	100	102 380	100	76.0

Of the total number of copies sold during the years 2010–2012, over 65 per cent were open access titles. However, the average number of copies sold per title was lower, than for titles not published in open access. The

turnover associated with these sales will not be discussed in this paper. Below, other influences on monograph sales are highlighted.

3.3 Influences on monograph sales

On the basis of the expertise of employees of AUP, several other possible influences on sales were defined: commercial potential, frontlist and backlist, and language. Each influence will be discussed below.

3.3.1 Commercial potential

An important part of the publishing process is determining how well a title will sell. The publisher will take into account several properties of the book and predict the number of copies that will sell. This then informs the print run, the number of copies made available for sale. The print run of the titles under consideration ranges from zero – no copies are printed beforehand – to 5000. The average print run of the books available on open access is lower than the average print run of books that are not available in this way. The average print run for books on open access is 459, whereas the average print run for the other titles is 652 (Table 2). In other words, the expected sales of books that are not available on open access are over 140 per cent of the expected sales of books on open access.

Table 2. Mean print run

	Number of titles	Mean print run
On open access	378	458.8
Not on open access	135	652.2
Total	513	509.7

Using a mean print run for a set of 513 titles is rather a crude instrument, which hides the complex decisions made for each title. Furthermore, if print runs are declining, this may also influence the data. First of all, the decline in print runs is not very clear in my data. The average print runs per publication year range from 286 to 782, without any clear trend. Secondly, the statistical analysis will take into account the sales data of each individual book per year and not use the averages described here. However, these averages give us a first clue about the commercial expectations of

the titles that have been made available on open access, compared with the other titles.

3.3.2 Frontlist and backlist

Publishers refer to the titles published in the current year as the 'frontlist'; all other titles are referred to as the 'backlist'. Experience shows that sales in the first year are generally higher than sales figures in subsequent years. This is the case with the books in our data set. The average number of copies sold of books published in 2011 was twice the average number the next year. The same holds true for books published in 2010: the average sales in 2010 are almost 2.5 times higher than the average number of copies sold in 2011 (Table 3).

Table 3. Front and backlist sales

Publishing date	*Number of titles*	*Mean sales in 2010*	*Mean sales in 2011*	*Mean sales in 2012*
2009 and before	314	49.8	31.1	21.4
2010	69	355.9	145.4	95.1
2011	68	–	195.2	96.0
2012	62	–	–	150.5
Total	513			

We could also look at the average sales per year. This seems to reveal a downward trend: in 2010, the average number of copies sold was 104; in 2011, 73; and in 2012, 56. Nevertheless, three years' data cannot be used to make conclusions about long-term developments. Moreover, it is possible that the decline in sales is spread evenly over all titles, regardless of whether they are published on open access. The study is set up to answer the question of whether publishing on open access makes a difference, taking into account the differences in frontlist and backlist sales, and the average sales per year.

3.3.3 Language

The analysed books were either published in Dutch or in English. Dutch-language books are more likely to be sold only in Dutch-speaking countries, whereas English-language books may be sold globally. The differences in

potential markets may influence the number of copies sold. In the examined group of titles, 63 per cent are published in English (Table 4).

Table 4. Titles per language

	English	*Dutch*	*Total*
On open access	268	110	378
Not on open access	54	81	135
Total	322	191	513

3.4 Data and Results

Here I measure the effect of the four influences – open access publishing, commercial potential, front and backlist, language – on sales. The ANOVA statistical method (analysis of variance) is used to check whether each influence has a significant effect. As a second step, the influences are combined to see in what way the interaction of these influences affects the number of books sold. The data are summarized in Table 5.

Table 5. Data: mean sales per influence

Influence		Mean sales	N	Percentage total sales
Open access publishing	With	66.28	1014	65.6
	Without	105.62	333	34.4
Commercial potential	Print run: 0	27.76	430	11.7
	Print run: 1–1000	69.62	782	53.2
	Print run: 1001–2000	217.1	105	22.3
	Print run: 2001–3000	454	24	10.6
	Print run: 3001–4000	324	3	0.9
	Print run: 4001–5000	446	3	1.3
Front and backlist	Frontlist	236.94	199	46.1
	Backlist	48.11	1148	53.9
Language	English	57.36	829	46.4
	Dutch	105.85	518	53.6

3.4.1 Separate influences

The effect of each influence is measured using the ANOVA procedure. This tests whether the differences among the mean sales of the books can be explained by chance. The results of each individual test are summarized in Appendix 1: ANOVA results per influence.

It is clear from the results that each influence by itself correlates with monograph sales in our data. So, while it is true that open access publishing is connected to sales, this is also true for commercial potential, front and backlist, and language. We can use two parameters to estimate the size of the effect: ω^2 and *F*-ratio. The first indicates that both commercial potential and front and backlist sales have a more profound effect on the number of copies sold than do open access or language. If we take into account the *F*-ratio, there is one outlier: front and backlist sales, whose *F*-ratio is 51.016, almost three times larger than the second-highest *F*-ratio. The difference between the mean sales of the frontlist and the mean sales of the backlist also reflect this large effect.

3.4.2 Combining influences

We did see that each influence is statistically significant, and this makes it harder to single out the effects of open access. It also became clear that there is a large difference between sales of the frontlist and sales from the backlist. The mean of all frontlist sales is almost 237, whereas the mean of all backlist sales is just over 48. In order to compensate for this large difference, the data are split into frontlist sales and backlist sales. Statistical methods are applied to these two data sets to measure the effect of open access publishing, combined with the influence of commercial potential and language.

3.4.3 Frontlist: data and results

The frontlist data consist of the sales data of 199 titles: the titles that were published in 2010, 2011, and 2012. Only the sales of the first year of publication are taken into account. At first glance, the difference is not very large between the mean sales of open access books (240.15) and those of books not available on open access (227.88). This contrasts strongly with language, where the mean sales of English-language books is approximately 40 per cent of the mean sales of books published in Dutch. The effects of commercial potential are visible: books with a higher print run did sell better

on average. A multifactor ANOVA procedure is used to test the effect of the combined influences.

When we look at the statistical analysis in Appendix 2: Frontlist results, the results of the frontlist can be explained by a combination of commercial potential and language. Open Access publishing does not have an effect in this situation. When we look at the effect size – measured by partial η^2 – of both print run and language, it becomes clear that commercial potential (measured by print run) plays the largest role. Of course, this is hardly surprising.

3.4.4 Backlist: data and results

The amount of available data from the backlist is much larger. Firstly, it contains the data of the 314 books published in 2009 and earlier which were sold during 2010–2012. On top of that, it contains the sales data for the years 2011 to 2012 of the 69 books published in 2010. And, lastly, the 2012 sales of the 68 books published in 2011 are also part of this set. The mean sales of the backlist are much lower: the backlist sales are on average 21 per cent of the frontlist sales mean. Compared with frontlist sales, the difference between mean sales of books on open access and mean sales of books not on open access is much larger: 82 for titles not on open access versus just under 37 for open access books. Still, the total number sold of backlist books not on open access is roughly 70 per cent of the number of backlist open access books sold.

Using the same procedure as before, the results for the backlist can be explained by a combination of commercial potential and open access. Language does not play a significant role. The results are listed in Appendix 3: Backlist results. Still, to get meaningful results from a multifactor ANOVA procedure, several preconditions must be met. The most important precondition is homogeneity of variance. In other words, the means used in the procedure should be evenly spread. The backlist data did not meet this condition, and so we must interpret the results with caution.

As a possible solution to overcome the statistical problems, the data can be split into smaller samples based on 'print run groups'. This creates four subsets of books with the same commercial potential, where each subset contains books that are published on open access and books that are not. Creating these subsets enables us to measure the effect of open access while controlling for the effects of commercial expectations. As a consequence, the subsets contain fewer data; this is most noticeable with the set 'print run 2001–3000', where the number of data items is as low as 20 ($N = 20$). For

each set, a one-way ANOVA procedure is performed to test the influence on open access on sales.

Table 6 presents the mean sales of books in the backlist, sorted by print run. The most striking difference between open access books and books not on OA can be found for print runs between 1001 and 2000. In this relatively small group (N = 92), the mean sale of titles not on open access is 201.94, compared with 68.21 for titles published on open access. Consistent with the discussion in the introduction, most of the titles published have small print runs: 1000 or less. In the group of titles with a print run of zero and the group of titles with a print run between 1 and 1000, the difference in mean sales between open access books and books not on open access is smaller.

Table 6. Backlist data: commercial potential and open access

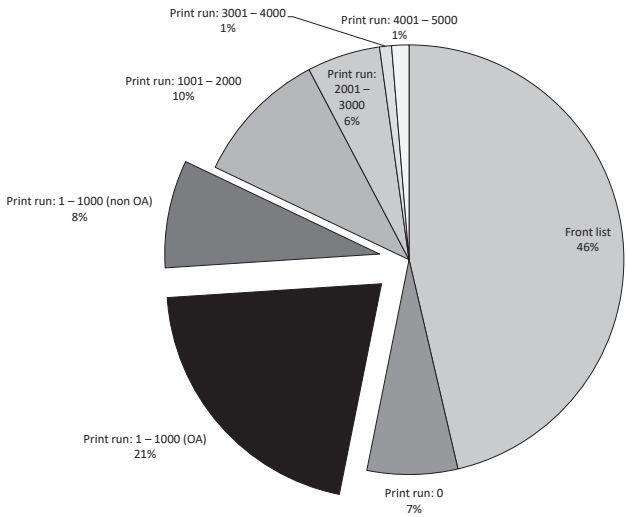
Commercial potential		Mean sales	N	Percentage total sales
Print run: 0	On open access	17.08	290	4.8
	Not on open access	21.65	95	2.0
Print run: 1–1000	On open access	41.84	509	20.8
	Not on open access	62.38	136	8.3
Print run: 1001–2000	On open access	68.21	61	4.1
	Not on open access	202.94	31	6.1
Print run: 2001–3000	On open access	213.57	7	1.5
	Not on open access	321.38	13	4.1
Print run: 3001–4000	Not on open access	324	3	0.9
Print run: 4001–5000	Not on open access	446	3	1.3

When the sales data are analysed using statistical methods, we see that open access publishing is a significant – negative – influence on the average number of copies sold in certain cases only: the subsets of books whose print run is between 1 and 1000 or between 1001 and 2000. No significant effect on books with a print run of zero or between 2001 and 3000 could be measured. Furthermore, the measured effect of open access on sales is much higher for the books with a print run between 1001 and 2000 than for the books with a print run between 1 and 1001. The results are fully described in Appendix 3: Backlist results.

At first glance, the outcomes of this paper run counter to the results of Snijder (2010) and OAPEN-NL (Ferwerda *et al.*, 2013). Here we see that making books available on open access has affected sales in certain circumstances. However, when we look at the total number of copies sold, the effects are not as strong as might be expected. This is best explained using the following illustrations.

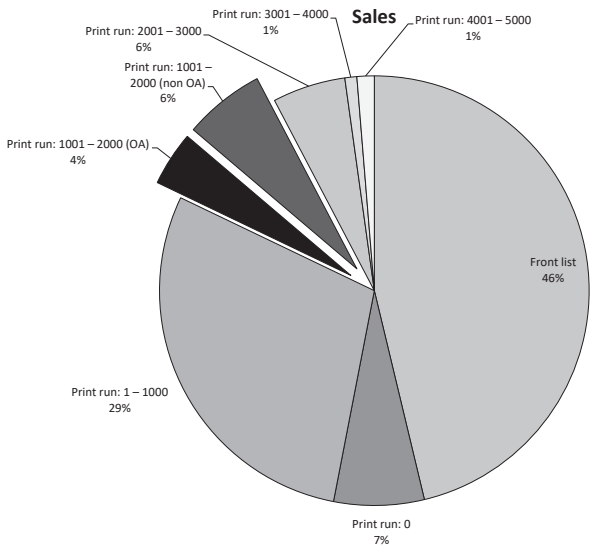
In Figure 1, the backlist sales of titles with a print run between 1 and 1000 are highlighted. The mean sales of books published in closed access are higher than those of books published in open access. However, the total number of copies sold is much lower: it amounts to 8 per cent of all sales. In contrast, the number of copies sold of titles under open access is 21 per cent of all sales.

Figure 1. Backlist sales, print run 1–1000.



The backlist sales of titles with a print run between 1001 and 2000 are highlighted in Figure 2. Here, the difference between mean sales of open access titles and those of titles not on open access is quite large: 68 versus 202. This does not lead to an equally large difference in total sales. We can see that 10 per cent of all copies sold are have a print run between 1001 and 2000. Here, four per cent are published on open access and six per cent that are not made available in this way.

Figure 2. Backlist sales, print run 1001–2000.



3.5 Discussion

Sales of monographs are influenced by several factors, and this paper tries to measure the effects of open access publishing, commercial potential, front and backlist sales, and language. The results show that all factors are influential, but the strengths of the measured effect are not equal. The difference between front and backlist sales is by far the greatest, and it was necessary to split the data to find meaningful results for the other influences.

The data used here did not come from a carefully set up experiment. Instead there was a bias in commercial expectations: the mean print run of the open access titles was 70 per cent of that of the books not published on open access. The difference is reflected in the mean sales of the two groups: the mean sales of open access titles are over 63 per cent of the mean sales of books not on open access. We might conclude that the lower commercial expectations of OA books is reflected in the mean sales. However, the number of open access titles available is larger, and the *total* number of copies sold is also larger: over 65 per cent of all sales.

The main question to answer is whether open access publishing is affecting the sales of monographs and how it compares with other influences

on monograph sales. The results for frontlist sales are clear: no significant effect of open access on sales could be found, after controlling for the effect of print run and language.

The answer for backlist sales is a bit more nuanced. Owing to limitations of the data, it is not possible to run the same procedure as used for the frontlist. Instead the backlist titles were split into four groups based on the commercial expectations. Taking into account this division, we can conclude that open access publishing has no effect on books with a print run of zero or between 2001 and 3000. Moreover, where an effect could be found, the effect size is quite different for categories of print run: a small effect for books with a print run between one and 1000 and a large effect for books with a print run between 2001 and 3000.

In the subsets of books with a print run higher than 0 but below 2000, open access has a negative effect. Nevertheless, the group of books with a print run between one and 1000 is very heavily skewed towards open access books: the data analysed contain over 3.6 times more data items ($N = 509$) for titles available under OA than data items for closed access titles ($N = 136$). The difference between the mean sales is much less: the mean sales of books not on open access is almost 1.5 times the mean sales of open access books in this group. This amounts to a much higher sale of copies of books on open access: almost 21 per cent of the total sales, compared with just over eight per cent for books not on open access. In an economic sense, the negative effect is not very important. The differences in the group of books with a print run between 1001 and 2000 are much more dramatic. But in this case the groups are small (with open access: $N = 61$; without open access: $N = 31$) and the number of books sold is 10 per cent of all sales. It is likely that the influence on revenue will not be very large.

From these data, it is not clear why sales of paper monographs are so lightly affected by free online versions. For a possible answer, we could look again at Snijder (2010) and OAPEN-NL (Ferwerda *et al.*, 2013; Snijder, 2010). There, academic libraries are described as a major purchaser of monographs. As long as availability on open access is not taken into account when paper monographs are acquired, the same pattern keeps emerging. However, we also saw that sales of monographs – whether available on open access or not – are far from soaring. Lack of budget at academic libraries is probably a major factor, as illustrated in ("Association of Research Libraries (ARL) :: ARL Statistics 2009-10," 2012).

The introduction to this paper discussed the sustainability of the current monograph publication system. From the results in the paper, we can conclude that using a hybrid model or closed access only does not make a large

difference to sales. The model under research does not lead to more sales of open access monographs, and the loss of sales is negligible. The data suggest that a hybrid model in which open access versions are made available in combination with paid-for print versions does not change the status quo. If the status quo is considered to be a broken publication system, a hybrid model is not an option to change things for the better. However, publishers who do well from selling paper monographs could consider making their titles available on open access as a way to enlarge the number of readers. Publishers who are making losses on monographs may want to change their business model in a more radical way than adopting a hybrid model.

3.6 Limitations

The data set used in this paper is large: it contains the sales data of 513 titles sold over a period of three years. Results from a large data set are less prone to be influenced by outliers, which helps to validate the outcomes. Nevertheless, the results are derived from one publisher and this makes it hard to establish whether other aspects – such as reputation or marketing budget – have influenced the results. Owing to the properties of the used sales data, the analysis was carried out on smaller subsets. Further research could establish whether the role of open access publishing in a hybrid model really is so small.

3.7 Acknowledgements

The author would like to thank Managing Director Jan-Peter Wissink of AUP for facilitating the research, and Professor Paul Wouters and Ludo Waltman of the Centre for Science and Technology Studies (CWTS) for commenting on a draft version of the paper.

3.8 Appendix 1: ANOVA results per influence

Table A1.1. ANOVA results per influence

Influence	*Results*
Open access	There was a significant effect of open access on monograph sales, $F(1, 529.828) = 10.974$, $p = 0.001$, $\omega^2 = 0.01$
Commercial potential	There was a significant effect of print run on monograph sales, $F(5, 11.083) = 16.727$, $p < 0.001$, $\omega^2 = 0.16$
Front and backlist	There was a significant effect of front and backlist on monograph sales, $F(1, 202.781) = 51.016$, $p < 0.001$, $\omega^2 = 0.14$
Language	There was a significant effect of language on monograph sales, $F(1, 660.003) = 17.216$, $p < 0.001$, $\omega^2 = 0.02$

Note: The assumption of homogeneity of variance was violated; therefore, the Welch *F*-ratio is reported for 'open access'; 'commercial potential'; 'front and backlist'; 'language'.

3.9 Appendix 2: Frontlist results

Table A2.1. Frontlist data: mean sales[Q31]

Influence		Mean sales	N	Percentage of total sales
Open access publishing	With	240.15	147	34.5
	Without	227.88	52	11.6
Commercial potential	Print run: 0	109.49	45	4.8
	Print run: 1–1000	179.99	137	24.1
	Print run: 1001–2000	949.46	13	12.1
	Print run: 2001–3000	1305.75	4	5.1
Language	English	158.14	134	20.7
	Dutch	399.4	65	25.4

Multifactor ANOVA

The results convey that the covariate print run was significantly related to sales, $F(1, 195) = 81.651$, $p < 0.001$, partial $\eta^2 = 0.295$. Also, the covariate language was significantly related to sales, $F(1, 195) = 22.577$, $p < .001$, partial $\eta^2 = 0.104$. However, no significant effect of open access on sales could be found after controlling for the effect of print run and language, $F(1, 195) = 2.83$, $p = 0.094$, partial $\eta^2 = 0.014$.

3.10 Appendix 3: Backlist results

Table A3.1. Backlist data: mean sales

Influence		Mean sales	N	Percentage total sales
Open access publishing	With	36.8	867	31.2
	Without	82	281	22.5
Commercial potential	Print run: 0	18.21	385	6.8
	Print run: 1–1000	46.17	645	29.1
	Print run: 1001–2000	113.61	92	10.2
	Print run: 2001–3000	283.65	20	5.5
	Print run: 3001–4000	324	3	0.9
	Print run: 4001–5000	446	3	1.3
Language	English	37.93	695	25.7
	Dutch	63.73	453	28.2

Multifactor ANOVA

Using the same procedure as deployed on the frontlist leads to the following result: the covariate print run was significantly related to sales, $F(1, 1144) = 234.618$, $p < 0.001$, partial $\eta^2 = 0.17$. However, the covariate language was not significantly related to sales, $F(1, 1144) = 2.17$, $p = 0.141$, partial $\eta^2 = 0.002$. Open Access has a significant effect on sales after controlling for the effect of print run and language, $F(1,1144) = 27.948$, $p < 0.001$, partial $\eta^2 = 0.024$.

Table A3.2. Backlist data: commercial potential

Commercial potential	Results
Print run: 0	No significant effect of open access on monograph sales could be found, $F(1, 126.225) = 1.25$, $p = 0.291$, $\omega^2 = 0.00$
Print run: 1–1000	Open access had a significant negative effect on monograph sales, $F(1, 179.348) = 7.364$, $p = 0.007$, $\omega^2 = 0.01$
Print run: 1001–2000	Open access had a significant negative effect on monograph sales, $F(1, 36.510) = 9.795$, $p = 0.003$, $\omega^2 = 0.13$
Print run: 2001–3000	No significant effect of open access on monograph sales could be found, $F(1, 18) = 0.449$, $p = 0.511$, $\omega^2 = 0.00$

Note: The assumption of homogeneity of variance was violated for 'Print run: 0', 'Print run: 1–1000', and 'Print run: 1001–2000'; therefore, the Welch *F*-ratio is reported.

4 Modes of access : The influence of dissemination channels on the use of open access monographs

Snijder, R. (2014). Modes of access: the influence of dissemination channels on the use of open access monographs. Information Research, 19(3), 166–183. Retrieved from http://www.informationr.net/ir/19-3/paper638.html

4.1 Introduction

Open access is much debated and in recent years has gained much attention in the literature. The scientific and scholarly impact of papers has been discussed extensively, for instance by Antelman (2004), who finds that freely published papers receive more citations – across a number of disciplines. Podlubny (2005) takes the citation analysis a step further and proposes a normalisation procedure, aimed at comparing the impact of scientists from different fields. Bollen *et al.* go beyond citations and investigate 39 impact measures, and conclude that "usage-based measures" may be a better indication of scientific impact (Bollen, Van de Sompel, Hagberg, & Chute, 2009).

Not only is the impact hotly debated but the economic aspects have also received much attention. A major discussion point is the merits of publishing a free version of a paper next to the 'official' version in a journal which is not freely accessible (green open access), versus the merits of directly publishing in an open access journal (gold open access) (Harnad *et al.*, 2004, 2008). Recently, the report *Accessibility, Sustainability, Excellence: how to expand access to research publications* by Finch *et al.* was heavily discussed (Finch *et al.*, 2013).

The discussion on the effects of open access on monographs does not attract the same amount of attention so far, and the amount of available research is small. Apart from running the OAPEN Library, the OAPEN foundation is currently involved in two pilot projects in the Netherlands ("OAPEN.nl website - English," n.d.) and the UK ("JISC - OAPEN-UK," n.d.) experimenting with open access monograph publishing. The first results of the OAPEN-UK pilot are discussed by Collins & Milloy (2012). In September

2013, the results of the Dutch pilot project were published (Ferwerda *et al.*, 2013).

4.2 Dissemination channels

This paper will focus on a different aspect: dissemination channels. In the literature on open access, dissemination channels seem to be a given. If it is discussed at all, dissemination is described as making papers available in an institutional repository. This paper is the first to analyse the effects of several dissemination channels in an open access environment.

Here we examine the monograph downloads of the OAPEN Library, which was officially launched in September 2010 (OAPEN Consortium, 2011; Open Access Publishing in European Networks, 2010b). It is a Web based collection of monographs, mainly in the field of Humanities and Social Sciences (HSS). All books are available in open access and users can search the Website in several ways. Each book also has a unique Web address and can be downloaded directly without searching the Website. These addresses – combined with metadata describing the books – are made available on the OAPEN Website and through several aggregators. This is described in more detail in Snijder (2013a).

This paper examines the download data of the OAPEN Library, which was gathered during a period of six months. The data consists of the number of downloads per month per provider. Here we define a provider as the organisation that grants the user access to the internet. Furthermore, the data contains information on whether a book was downloaded via the OAPEN Website or directly. Because the data were aggregated monthly, we can distinguish three situations: firstly, a book was downloaded a certain number of times through a provider via the Website only; secondly, a book was downloaded a certain number of times through a provider using the direct download address of the book; thirdly, a book was downloaded a certain number of times through a provider via the Website and also a certain number of times directly. In the last case, the readers related to that provider use a combination of ways to access the book.

It is not unreasonable to assume that each provider caters for several people. In the case where all readers only use the Website or only use direct downloads, their preference seems to be aligned. If – in the same month – a portion of the readers use the Website and another portion of the readers prefer direct downloads, this may hint at another 'group configuration'. In this case, other aspects of usage could also differ, which is why this is

analysed separately. Thus, the download data stems from three channels: Website only, Website and direct access; and direct access only.

As the data is available through several channels, it may be useful to investigate the literature on multichannel management. This field looks at the challenges that retailers face in the deployment of multiple channels to reach their customers. While typical research in this field looks at the differences between offline channels such as stores and online channels such as Websites, parts of the theoretical framework could be applied to this paper.

The multichannel management framework is based on theories on the adoption of innovations, explaining if and why people will use new channels. On this layer, the specific aspects of working with multiple (retail) channels are discussed. According to Rogers (1995), several factors influence the use of innovations: the relative advantage of the innovation, its fit with existing usage patterns, the perceived complexity, the ability to try out the innovation, the perceived risk related to adoption, and the degree to which adoption and use can be observed by others (Rogers, 1995).

The work of Rogers is paired to the technology adoption model (TAM) and its extension TAM2. TAM states that perceived usefulness and perceived ease of use are drivers of innovation adoption; TAM2 extends this framework to social influence processes (subjective norm, voluntariness, and image) and cognitive instrumental processes (job relevance, output quality, result demonstrability, and perceived ease of use) (Davis, 1989; Davis, Bagozzi, & Warshaw, 1989; Venkatesh & Davis, 2000). Neslin *et al.* identified five "key challenges" in multichannel management: data integration, understanding customer behaviour, channel evaluation, allocating resources across channels and coordinating channel strategies. In a later paper, the list of relevant aspects has grown to thirteen (Neslin *et al.*, 2006; Neslin & Shankar, 2009). Basically, the questions revolve around whether or not to deploy a multichannel strategy, how to set up different channels, and how to evaluate the results.

What aspects of multichannel management can be used here? Instead of offline versus online channels, we are discussing different online channels. We envision different users with different needs. They are not paying customers, and researching and 'purchasing' in an open access environment are more or less the same action. Searching for information in the field of humanities and social sciences is covered by many authors. Shen discusses the many channels used by social scientists, grouping them in internal and external electronic and paper resources, combined with "external human resources" Shen (2007, p.8). Bulger *et al.* discuss humanities

scholar's search behaviour through six use cases where scholars employed a range of resources and technologies (Bulger *et al.*, 2011). Wang *et al.* use an international angle by discussing the scholars in the USA, Greece and China (Wang, Dervos, Zhang, & Wu, 2007). Griffiths and Brophy focus on students' online search behaviour, and describe the strong preference for search engines – especially Google – compared to the library catalogue or other sources (Griffiths & Brophy, 2005). Lamothe discusses the growing usage of e-books in an academic library (Lamothe, 2010).

Channel evaluation also has implications for resource management: the results help to decide where to invest the most time and money. This goes beyond managing IT systems, it also affects marketing decisions. In short, multichannel management aims to create an optimal strategy in a given environment.

If we combine search behaviour with the decision to use a specific channel, we arrive at the following research question: does the usage based on the channel "Website only" differ from usage based on "direct access only" or from usage from a combination of those channels? The answer has implications for open access publishing as it may help to optimise the dissemination of open access monographs.

First, the download data is analysed quantitatively: counting the number of downloads per channel. Then, the qualitative analysis tries to find an answer to the question of whether properties of the users, their infrastructure or the properties of the book themselves have a significant impact on the usage per channel.

4.3 Quantitative analysis

In this section, the data set is described, followed by the number of downloads per channel. The number of monograph downloads is an indication of readership. Whilst we can assume that the more a monograph has been downloaded, the more it has been read we cannot, however, state that 100 downloads equal equates to 100 people reading the monograph cover to cover.

4.3.1 The data set

The data set consists of the download data of 979 books, published by thirty-five different publishers. The books are published in ten different languages. By far the largest number of the downloaded books are in English. The 979

monographs in the data set were downloaded 152,662 times in the first six months of 2012.

Table 1 Languages: number of titles

Language	Number of titles	Percentage
English	514	52.5%
German	164	16.8%
Dutch	125	12.8%
Other languages	176	18.0%
Total	979	100%

The ratios of the downloads per language are more or less in line with the percentages of published languages. This is discussed in more detail in the qualitative analysis. Annex 2 contains the complete list of languages.

Table 2 Languages: number of downloads

Language	Downloads	Download percentage
English	8,8003	57.6%
German	3,2632	21.4%
Dutch	1,9025	12.5%
Other languages	1,3002	8.5%
Total	152,662	100%

The following table lists the ten most downloaded subjects. This is a fraction of all available subjects: the complete data set contains eighty-three different subjects. The classification used is the BIC standard subject categories (BIC) (Book Industry Communication, 2010). The question of whether language or subject has a measurable influence on channel usage will be discussed in the qualitative analysis.

Table 3 Subjects: most downloaded

Subject	Number of titles	Percentage
History (HB)	165	17.0%
Politics & government (JP)	148	15.3%
Society & culture: general (JF)	80	8.2%
Sociology & anthropology (JH)	62	6.4%
Film, TV & radio (AP)	32	3.3%
Literature: history & criticism (DS)	37	3.8%

Subject	Number of titles	Percentage
Philosophy (HP)	25	2.6%
Religion & beliefs (HR)	23	2.4%
Science: general issues (PD)	34	3.5%
Laws of Specific jurisdictions (LN)	32	3.3%
Other subjects	332	34.2%
Total	979	100%

As before, the ratios of the downloads per subject are more or less in line with the percentages of published subjects. This is discussed in more detail in the qualitative analysis. Annex 3 contains the complete list of subjects.

Subject	Number of downloads	Download percentage
History (HB)	23,624	15.5%
Politics & government (JP)	19,167	12.6%
Society & culture: general (JF)	13,520	8.9%
Sociology & anthropology (JH)	9,033	5.9%
Film, TV & radio (AP)	6,571	4.3%
Literature: history & criticism (DS)	6,786	4.4%
Philosophy (HP)	5,896	3.9%
Religion & beliefs (HR)	4,506	3.0%
Science: general issues (PD)	3,796	2.5%
Laws of Specific jurisdictions (LN)	7,002	4.6%
Other subjects	52,761	34.6%
Total	152,662	100%

We saw that the 979 books were downloaded 152,662 times in the first six months of 2012. The books were accessed through 6176 different providers which are based in 166 countries. We stated before that a provider is defined as the organisation that grants the user access to the internet. In some cases, the provider is an organisation such as a university or a government agency. In other cases, this is an Internet Service Provider (ISP), such as Comcast in the USA or Ziggo in the Netherlands. The providers will be discussed in more detail in the qualitative analysis.

4.3.2 Downloads per dissemination channel

The downloads were measured per provider per channel per month. So, if a provider downloaded the same monograph more than once in the same month, using the same channel, the number of downloads were added. In some instances, a provider downloaded a monograph several times a month via the Website and also via direct access. In those cases, the downloads were added to the combined channel "Website and direct access". In other instances, a monograph was only downloaded via the Website, or the monograph was only downloaded through direct access only. Then the downloads were added to the channels "Website only" or "direct access only" respectively.

Using this procedure, the following data becomes available:

Channel	Number of downloads	Percentage
Website only	11,546	8%
Website and direct access	29,453	19%
Direct access only	111,663	73%
Total	152,662	100%

The data shows that usage is dominated by direct access only. This implies that almost three quarters of all downloads come from users who do not use the Website http://www.oapen.org/, but find the books via other means. This kind of usage is made possible by making the metadata[12] of the books – including a direct download URL – directly available to all interested parties, including libraries and content aggregators. The channel "Website and direct access" contains a combination of downloads via the Website and direct access. Here again, the portion of downloads via direct access is larger than the downloads via the Website. It is clear that most readers find the books through other routes than the OAPEN Library Website.

The usage data revealed that 24% of the visits to the OAPEN Library Website lead to downloading one or more titles. However, this percentage cannot be compared to the usage data of other systems. If 100 OAPEN monographs were downloaded via a library catalogue, how many searches were conducted which did not result in a download taking place? Therefore, we do not know whether the OAPEN Library Website is a more efficient way to search compared to other systems.

12 The metadata is licensed under a Creative Commons Zero licence, which makes it free to use under any circumstance.

We discussed before that multichannel management aims to create an optimal strategy in a given environment. The goal of open access publishing is to remove barriers to access, and it makes sense to investigate how to maximize the dissemination of open access monographs. We saw that the direct access channel is far more used than the other channels and this has serious consequences for managing and optimising the service: from a dissemination point of view it makes more sense to invest in metadata and the dissemination of metadata then to spend resources on the Website. It is important that any system used for open access dissemination is capable of exporting metadata in formats that can be used by content aggregators or the systems used by prospective readers. Apart from library catalogues, search engines may be a much-used research tool, and investing resources in optimal coverage by the likes of Google and Bing may be beneficial.

4.4 Qualitative analysis

The goal of the qualitative analysis is to establish whether user's characteristics (i.e. their infrastructure) or the collection are influential factors on channel usage. Firstly, user characteristics are discussed. The download percentages of the quantitative analysis are used as a benchmark, and are compared to the actual values found using an independent t-test. A factor is considered influential if the difference between the usage numbers is statistically significant and the effect size is not small.

4.4.1 Characteristics of users and dissemination channels

Readers are placed in several groups: academic; government; business; non-profit organisations and the general public. While academic users could be seen as the main audience for monographs, readers of other backgrounds have equal access to the monographs in the OAPEN Library. The users are categorised based on the data from the OAPEN logs, combined with public data.

The OAPEN Library is a Web based service, and its logs contain the Web address of the providers. So, if researchers at Leiden University download a book using their office equipment, the Web address (www.leidenuniv.nl) of that university will be logged. Basic information such as address and telephone number are publicly available and can be found using the so called 'WHOIS protocol' ("WHOIS - Wikipedia," n.d.). By combining the usage data and information about the provider, we can make assumptions about who is downloading a specific monograph.

A large portion of the providers are not universities or government agencies, but internet service providers (ISP). If the provider is an ISP, the user cannot be linked to an organisation. We cannot assume that all usage through an ISP comes from people browsing the internet at home. If the internet infrastructure in a certain country is highly developed, chances are that each organisation is capable of giving direct internet access to their members. If the internet infrastructure is less well developed, a large portion of the organisations in that country do not directly provide internet access but rely on the services of an Internet Service Provider.

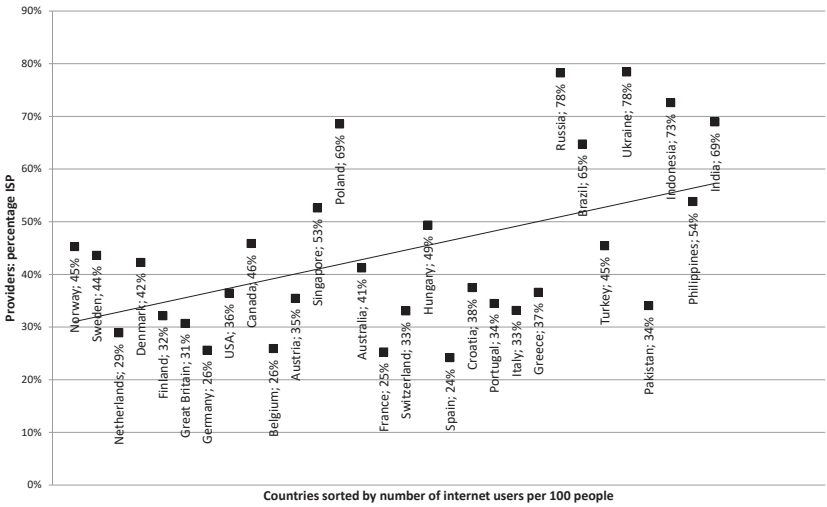
Of course, it is always possible that 'ISP users' from a country with a highly-developed internet infrastructure are in fact academics working from home after office hours. The available data does not contain the (local) time of the download, which makes determining whether a reader is downloading during office hours impossible. Furthermore, if the reader is not acting in a professional capacity, the chances are also higher that the download started after office hours. The difference in access to scholarly and scientific literature for academics compared to others is quite large; using the credentials of the academic institution allows direct access to all kinds of literature behind pay walls. It might therefore be more efficient to use these credentials not only at the office, but also after office hours.

If we want to divide ISP usage in those two categories, we need a way to determine the state of a country's infrastructure. This is done by using a World Bank publication: *The Little Data Book on Information and Communication Technology 2011* (World Bank & Lewandowski, 2010). It lists several statistics per country, and one of them is the number of internet users per 100 people. If there is a connection between the state of the infrastructure and the percentage of downloads through ISPs, the percentage of 'ISP usage' is lower for highly developed internet infrastructures.

This assumption was tested by charting the measured downloads from thirty countries and the percentage of ISP usage. The found values were set against the amount of internet users per 100 people. Because the country of each provider is known, it was possible to select the countries with the highest number of downloads. The selected thirty countries are responsible for almost 92% of all downloads.

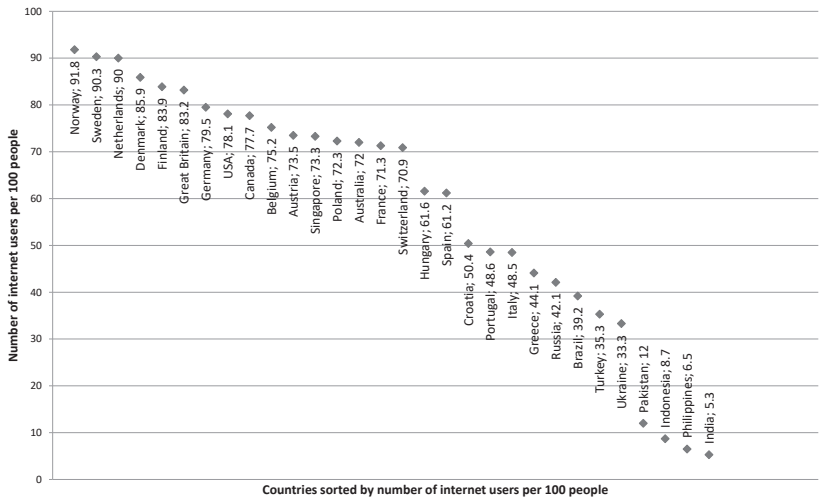
The first chart depicts the percentage of downloads via an ISP, sorted by the amount of internet users per 100 people. In this chart, we see that there is a trend toward a higher percentage of downloads via an ISP, when the number of internet users per 100 people decreases.

Figure 1 ISP percentage



The second chart depicts the number of internet users per 100 people. Here we see a decrease from 91.8 internet users per 100 people in Norway to 5.3 internet users per 100 people in India. Somewhere between these two extremes we need to set a cut-off point to determine which countries have a highly-developed internet infrastructure. Within these countries, the chances are higher that ISP usage from these countries is from "non-professional" users. This distinction is used in the qualitative analysis, to determine whether the internet infrastructure influences downloads via the different channels.

Figure 2 Internet users per 100 people



between Switzerland with 70.9 internet users and Hungary with 61.6 internet users per 100 people. Based on this, the threshold is set to seventy internet users per 100 people. Countries with seventy or more internet users per 100 people are considered to have a highly-developed infrastructure. The same threshold is also used in (Snijder, 2013b).

4.4.2 Type of users and dissemination channels

Now we can look at the download percentages of the different user groups. The number of downloads per channel differ wildly and because of that, there is a large difference in the absolute number of downloads per group. For instance, the number of downloads by academic readers through the "direct access only" channel is almost seven times the number of academic downloads through the Website only.

Is there a connection between user type and dissemination channel? Regardless of the channel, most of the usage comes from three groups: Academic, ISP and ISP high internet usage. As academics are the intended audience for monographs, it is not very surprising to see a large proportion of usage that originates from academic institutions. Furthermore, the academic community is large. As discussed before, it was not possible to determine whether the role of users in the group "ISP" was academic or otherwise. The members of the group "ISP – high internet" are more likely to be 'non-professional' users. Based on that we might conclude that

disseminating open access books helps to make scholarly content available to the public. In all channels, the usage by non-profit, government and business organisations is small, compared to academic and ISP related usage.

From the quantitative analysis it becomes clear that 8% of the usage comes from the channel Website only, 19% from the channel Website and direct access, and 73% through direct access only. We can use these percentages as a baseline for the expected downloads for each user group, and compare it to actual number of downloads per channel. Using the difference between those amounts – expressed as the percentage of the expected value – we find no significant effect for user type: $t(17) = -0.541$, $p = .595$. Based on the lack of significant differences on channel usage, we can conclude that the type of user plays a minimal role in channel usage.

Type of user	Website only		Website and direct access		Direct access only		All channels
	Actual	Expected	Actual	Expected	Actual	Expected	Actual
Academic	3,005	2,312	5,821	5,490	20,068	21,092	28,894
Business	49	393	11	933	4,849	3,583	4,909
Government	162	171	17	406	1,959	1,561	2,138
Non-profit	136	121	31	287	1,346	1,105	1,513
ISP - high internet usage	4,138	6,134	14,000	14,567	58,531	55,968	76,669
ISP	4,056	3,082	9,573	7,323	24,910	28,134	38,539
Total	11,546	12,213	29,453	29,006	111,663	111,443	152,662

4.4.3 Characteristics of internet infrastructure

Dividing the internet structure in highly developed and less well developed countries is not only useful to differentiate between user groups but is, in itself, also a possible influence on channel usage. We might expect that readers from countries with a highly-developed infrastructure have different download patterns compared to those with more limited bandwidth. Annex 1 lists the countries with highly developed infrastructure. When we look at overall usage – not taking into account the different channels – the difference between the two groups is clear: the number of downloads from countries with a highly-developed infrastructure is more than twice the number of downloads from the rest of the world.

The same percentages as before are used as a baseline for the expected downloads, and again those numbers are compared to the actual number of downloads per channel. Using the difference between those amounts – expressed as the percentage of the expected value – we find no significant

effect for internet infrastructure: $t(5) = -0.418$, $p = .639$. Based on the lack of significant differences on channel usage, we can conclude that internet infrastructure plays a minimal role in channel usage.

Internet infrastructure	Website only		Website and direct access		Direct access only		All channels
	Actual	Expected	*Actual*	*Expected*	*Actual*	*Expected*	*Actual*
Less than high usage	5,051	3,768	12,226	8,948	29,817	34,378	47,094
High usage	6,495	8,445	17,227	20,058	81,846	77,065	105,568
Total	11,546	12,213	29,453	29,006	111,663	111,443	152,662

4.4.4 Characteristics of content and dissemination channels

Is there a connection between characteristics of the content – the monographs – and dissemination channels? In this section, we examine two aspects: subject and language. Not all languages or subjects will be analysed: the three most downloaded languages and ten most downloaded subjects are examined.

4.4.5 Language and dissemination channels

It seems obvious that language influences the use of the monographs, as readers are unlikely to download a book in a language they cannot read. The high usage of monographs in the English language is directly visible, but we have to take into account the large number of books available in that language. The question is whether language usage differs significantly from expected values.

In the description of the data set, we saw that 52.6% of the books were written in English, 16.7% in German and 12.9% in Dutch. If we apply these percentages to the number of downloads per dissemination channel, we can compute the expected values. Using the difference between those amounts – expressed as the percentage of the expected value – we find no significant effect for language: $t(11) = -1.229$, $p = .245$. Based on the lack of significant differences on channel usage, we can conclude that language of the monographs does not play a role in channel usage.

Language	Website only		Website and direct access		Direct access only		All channels
	Actual	Expected	Actual	Expected	Actual	Expected	Actual
English	9,808	6,073	20,389	15,492	57,806	58,735	88,003
German	471	1,928	4,472	4,919	29,318	18,648	32,632
Dutch	396	1,489	2,843	3,799	14,157	14,405	19,025
Other languages	871	2,055	1,749	5,243	10,382	19,876	13,002
Total	11,546	11,546	29,453	29,453	111,663	111,663	152,662

Still, the percentage of downloads of English language books via the Website is relative high, and this raises the question of whether users primarily search using English terms. To test this, a small sample was analysed. Of all queries in one month, a list was created of searches that occurred at least twice. This created a set of 2,219 different queries.

	Number of queries	Percentage
In English	1,074	48.4%
Not in English	1,145	51.6%
Total	2,219	100%

The percentage of 'non-English' queries was more that 51%. Nevertheless, this group also contained search terms that exist not only in the English language, but also in Dutch and German. If we analyse this group, five 'ambiguous' terms account for more than 62% of queries: "film"; "water"; "IMISCOE"; "Iran"; "Islam". So, a large percentage of all the examined queries are at least 'compatible' to English. It is therefore safe to assume that most searches are indeed in English, which would partly explain the results. The large amount of available English language books might be another factor.

Multilingual terms	Number of queries	Percentage
"film"	348	30.4%
"water"	167	14.6%
"IMISCOE"	150	13.1%
"Iran"	31	2.7%
"Islam"	24	2.1%
Other terms	425	37.1%
Total	1,145	100%

4.4.6 Subject and dissemination channels

The last aspect to analyse is the subject of the monographs. Are the users of the OAPEN Library interested in certain subjects or do the download patterns closely follow the spread of subjects amongst the books? We have found the percentages of titles with a certain subject in the quantitative analysis. The expected number of downloads per channel are computed by applying these percentages to the number of books downloaded per channel, and the actual number of downloads are compared against the benchmark values. Using the difference between those amounts – expressed as the percentage of the expected value – we find no significant effect for subject: $t(32) = 1.507$, $p = .142$. Based on the lack of significant differences on channel usage, we can conclude that subject does not play a role in channel usage.

Subject	Website only		Website and direct access		Direct access only		All channels
	Actual	*Expected*	*Actual*	*Expected*	*Actual*	*Expected*	*Actual*
History (HB)	1,751	1,963	5,647	5,007	16,226	18,983	23,624
Politics & government (JP)	1,649	1,767	3,713	4,506	13,805	17,084	19,167
Society & culture: general (JF)	1,512	947	2,645	2,415	9,363	9,156	13,520
Sociology & anthropology (JH)	1,061	739	1,900	1,885	6,072	7,146	9,033
Film TV & radio (AP)	869	381	1,422	972	4,280	3,685	6,571
Literature: history & criticism (DS)	324	439	1,340	1,119	5,122	4,243	6,786
Philosophy (HP)	304	300	1,413	766	4,179	2,903	5,896
Religion & beliefs (HR)	367	277	879	707	3,260	2,680	4,506
Science: general issues (PD)	505	404	605	1,031	2,686	3,908	3,796
Laws of Specific jurisdictions (LN)	134	381	485	972	6,383	3,685	7,002
Other subjects	3,070	3,949	9,404	10,073	40,287	38,189	52,761
Total	11,546	11,546	29,453	29,453	111,663	111,663	152,662

However, when a dissemination channel is used more – the channel "direct access only" is used for of 73.1% of all downloads, while usage through the Website only is 7.6% – the number of subjects also grows. This is illustrated by the fact that the ten subjects listed here cover almost 74% of

all downloads occurring via the Website only. In contrast, the percentages drop for the other channels to 63.8%.

4.5 Conclusions

This paper is the first to analyse the effects of several dissemination channels in an open access environment. Its goal is to help determine an optimal strategy to achieve maximum distribution of open access monographs. The books are made available via the OAPEN Library Website, via direct downloads or a combination of those two. It is interesting to note that a large proportion of readers who directly download the monographs do not use the Website; they have found the description of the books via other means.

From the quantitative analysis, the dominance of one channel is clear. The data shows that 73% of all downloads occurred via the channel "direct download". This implies that almost three quarters of downloads come from users who do not use the Website http://www.oapen.org/, but find the books through other systems or Websites.

The qualitative analysis revealed that regardless of the channel, most usage comes from three groups: Academic, ISP and ISP high internet usage. Other user groups – Business, Government and Non-profit – are not highly represented. When looking at the usage per group, no effect on channel usage could be established. The internet infrastructure is another factor that was taken into account. While the digital divide between users from countries with a highly-developed internet infrastructure and users from less well-off countries is very clear, no effect on channel usage could be found. The same holds true for the aspects of the books themselves: the analysis could not find any effect on channel usage for both the language and subjects of the monographs.

The goal of multichannel analysis is to determine the optimal usage of resources: what configuration leads to the best results? The definition of "best results" in an open access environment differs from a commercial environment. The objective is not financial gain, but maximum dissemination. In the OAPEN Library, readers can access books via three channels. Firstly, the Website, which is optimised for search – it does not only contain metadata, but also enables full text search. Furthermore, it contains browsing functions as a means to enable serendipity. In contrast to this, the direct search channel functions in a different way. It is based on metadata only, which is incorporated into systems outside the OAPEN Library. Full text

search on the contents of the books is not possible. The third channel is a combination of both.

The results show that most readers are using the "direct download" channel, in spite of the fact that the OAPEN Library Website offers functions that are not available via other channels. A possible answer may be found in the theoretical models on the use of innovations discussed in the introduction. There we saw several factors influencing the use of new systems, such as its fit with existing usage patterns, perceived ease of use and social norms. It is possible that users of the "direct download" channel prefer their 'own' systems, which are familiar and are part of their routine and environment. In that case, learning to use a new interface may not be seen as a worthwhile investment. But who are the principal users of the OAPEN Library? The analysis revealed that current users are based in academic institutions or use an ISP. Users based in businesses, governmental or non-profit organisations are far less common. Also, the digital divide between upcoming countries and the 'developed' countries is a large factor: two-thirds of the downloads occurred from countries with a highly-developed internet infrastructure. And although the OAPEN Library contains books in German, Dutch, Italian and other languages, the majority of the books – and the majority of the readers – use English.

How does this compare to the goal of maximum dissemination? A recurring theme in the discussion on open access is making scientific and scholarly results available to members of academia who cannot access the information behind a pay wall. Seen from that perspective, the current situation is quite a success: academic institutions are responsible for a large portion of the downloads. However, when we look at other possible patrons, the picture is less rosy. In the collection of the OAPEN Library, the subjects politics & government, society & culture and sociology & anthropology are well covered. Those books may contain useful information for governmental organisations – for instance in the field of immigration studies, which is a much-debated topic in Europe and North America. Nevertheless, there is not much usage from governmental organisations, nor from non-profit organisations. Does the form – monographs – not fit within the informational habits of those potential users, or is the OAPEN collection not embedded in the information systems used?

When we compare the usage from countries with a highly-developed internet infrastructure to the usage from the rest of the world, the difference is striking. The first group of countries contains twenty-seven countries, yet it has downloaded twice as much as books. Here we see that making books freely available does not automatically take away other barriers to access.

The language of the publications may be another issue to research. More than half of the analysed books are written in English, and the download percentage of English language books is also roughly 50%. It is possible that the overall usage is at least partly shaped by the number of books available in a certain language. In other words, if the collection contained a larger percentage of monographs in another language – for instance French, Spanish and Portuguese – how might that affect the usage?

The results imply that making the metadata available in the user's systems – the infrastructure used on a daily basis – ensures the best results. So, to achieve the optimum amount of usage, firstly we must identify users who are not using the data; secondly, we have to understand how they search for information and thirdly we have to establish what is the best way to make our data available. Researching those questions would bring the goal of maximum dissemination a little closer. These challenges are not only faced by the OAPEN Foundation, but are shared by all organisations that disseminate open access publications or data.

4.6 Limitations

The data set used in this paper is large and rich: it contains the data of 979 monographs which were downloaded 152,662 times in the first six months of 2012. Several aspects of the monographs are described: language and subject. Furthermore, several characteristics of each download are available: the name of the provider and the channel used.

Nevertheless, as there is no authority data obtainable, the categorisation of the providers is not checked. Another aspect linked to the categorisation of providers is determining its country of origin, based on the available WHOIS information which always links one country to a provider. If an organisation operates in several countries – such as a NGO or a multinational – this will not be reflected in the data. Also, the subject information of the books has been simplified. These aspects may have had an influence on the qualitative analysis.

The timeframe could also be considered. The data was captured during a six-month period, and owing to the rapid pace of technological development on the internet, it would be interesting to compare the results with data from another period. Because this research is the first of its kind, no best practises have been established.

4.7 Acknowledgements

The author would like to thank Professor Paul Wouters of the Centre for Science and Technology Studies (CWTS) for commenting on the draft version of this paper.

4.8 Annex 1: list of countries with a highly-developed internet infrastructure

According to *The Little Data Book on Information and Communication Technology 2011*, the following countries have 70 or more internet users per 100 people:

- Andorra
- Australia
- Austria
- Belgium
- Bermuda
- Brunei
- Canada
- Denmark
- Estonia
- Finland
- France
- Germany
- Great Britain
- Iceland
- Japan
- Luxembourg
- Netherlands
- New Zealand
- Norway
- Poland
- Singapore
- Slovakia
- South Korea
- Sweden
- Switzerland
- United Arab Emirates
- USA

4.9 Annex 2: downloads per language

Language	Downloads	Percentage
English	88,003	57.6%
German	32,632	21.4%
Dutch	19,025	12.5%
Italian	8,586	5.6%
Danish	1,387	0.9%
French	629	0.4%
English, Latin	594	0.4%
German, Latin	488	0.3%
Spanish	476	0.3%
French, Latin	395	0.3%
German; English	236	0.2%
Norwegian	115	0.1%
Welsh	96	0.1%
	152,662	100%

4.10 Annex 3: downloads per subject

Subject	Downloads	Percentage
History (HB)	23,624	15.5%
Politics & government (JP)	19,167	12.6%
Society & culture: general (JF)	13,520	8.9%
Sociology & anthropology (JH)	9,033	5.9%
Film, TV & radio (AP)	6,571	4.3%
Literature: history & criticism (DS)	6,786	4.4%
Philosophy (HP)	5,896	3.9%
Religion & beliefs (HR)	4,506	3.0%
Science: general issues (PD)	3,796	2.5%
Laws of Specific jurisdictions (LN)	7,002	4.6%
History of art / art & design styles (AC)	4,092	2.7%
Humanities (H)	3,810	2.5%
Society & social sciences (J)	3,439	2.3%
linguistics (CF)	3,317	2.2%
Economics (KC)	3,118	2.0%
Literature & literary studies (D)	2,415	1.6%
Industry & industrial studies (KN)	2,252	1.5%
Business & management (KJ)	2,117	1.4%
The environment (RN)	1,848	1.2%
Law (L)	1,673	1.1%
Biology, life sciences (PS)	1,566	1.0%
Theatre studies (AN)	1,352	0.9%
Library & information sciences (GL)	1,350	0.9%
Interdisciplinary studies (GT)	1,348	0.9%
Architecture (AM)	1,334	0.9%
Archaeology (HD)	1,283	0.8%
Psychology (JM)	1,034	0.7%
Economics, finance, business & management (K)	956	0.6%
International law (LB)	907	0.6%
Education (JN)	891	0.6%
The arts: general issues (AB)	851	0.6%
Digital lifestyle (UD)	820	0.5%
Industrial chemistry & manufacturing technologies (TD)	650	0.4%
Music (AV)	600	0.4%
Educational material (YQ)	578	0.4%
Jurisprudence & general issues (LA)	531	0.3%
... (HJ)	487	0.3%
Poetry (DC)	391	0.3%
Social services & welfare, criminology (JK)	371	0.2%
Language (C)	361	0.2%
… (JR)	361	0.2%
Medicine (M)	358	0.2%

Subject	Downloads	Percentage
ELT background & reference material (EB)	328	0.2%
Warfare & defence (JW)	296	0.2%
Romance (FR)	287	0.2%
Agriculture & farming (TV)	285	0.2%
Prose: non-fiction (DN)	282	0.2%
Earth sciences (RB)	266	0.2%
Memoirs (BM)	248	0.2%
Biography & True Stories (B)	228	0.1%
Sports & outdoor recreation (WS)	227	0.1%
Civil engineering, surveying & building (TN)	224	0.1%
Finance & accounting (KF)	218	0.1%
… (HF)	204	0.1%
Adventure (FJ)	203	0.1%
Fiction & related items (F)	188	0.1%
… (QM)	185	0.1%
Graphical & digital media applications (UG)	182	0.1%
Encyclopaedias & reference works (GB)	181	0.1%
Medicine: general issues (MB)	172	0.1%
Databases (UN)	168	0.1%
Geography (RG)	161	0.1%
Mathematics (PB)	158	0.1%
Language teaching & learning (other than ELT) (CJ)	155	0.1%
Crime & mystery (FF)	144	0.1%
… (DV)	127	0.1%
Antiques & collectables (WC)	115	0.1%
Local interest, family history & nostalgia (WQ)	113	0.1%
Earth sciences, geography, environment, planning (R)	97	0.1%
Reference, information & interdisciplinary subjects (G)	89	0.1%
Art treatments & subjects (AG)	87	0.1%
Environmental science, engineering & technology (TQ)	87	0.1%
Modern & contemporary fiction (post c 1945) (FA)	79	0.1%
Other branches of medicine (MM)	77	0.1%
… (JB)	75	0.0%
Astronomy, space & time (PG)	68	0.0%
… (LK)	66	0.0%
Museology & heritage studies (GM)	59	0.0%
Biography: general (BG)	56	0.0%
Computer science (UY)	44	0.0%
Fiction: special features (FY)	39	0.0%
Art forms (AF)	24	0.0%
… (JS)	8	0.0%
	152,662	100%

5 Better sharing through licenses? : Measuring the influence of Creative Commons licenses on the usage of open access monographs

Snijder, R. (2015). Better Sharing Through Licenses? Measuring the Influence of Creative Commons Licenses on the Usage of Open Access Monographs. *Journal of Librarianship and Scholarly Communication*, 3(1), eP1187. https://doi.org/10.7710/2162-3309.1187

5.1 Introduction

Open access (OA) and content licenses are closely intertwined. The first Budapest open access Initiative declaration (Chan *et al.*, 2002) – widely seen as the official birth of the open access movement – does not explicitly state the need for a license, but the Berlin Declaration ("Berlin Declaration on Open Access to Knowledge in the Sciences and Humanities," 2003) and the Bethesda Statement (Suber *et al.*, 2003) pose two conditions: a license permitting distribution and reuse; and a deposit of the contents in a technically suitable manner. The goal of the open access movement is to disseminate scholarly and scientific knowledge as widely as possible, and using licenses to systematically remove the barriers created by copyright restrictions is an important tool.

One of the best-known licenses used to achieve this is the Creative Commons (CC) license. The Creative Commons organisation describes its set of licenses as a "simple, standardized way to give the public permission to share and use your creative work – on conditions of your choice." ("About - Creative Commons," n.d.) These licenses enable the copyright owner to allow certain types of usage – such as copying or modifying the contents – while constricting other forms of use, for instance by prohibiting commercial reuse. The six licenses vary in the amount of restrictions placed on the reuse of the work.

Beyond alerting individual users of their reuse rights, there is another important aspect to these licenses. Placing Creative Commons license code on digital content not only provides a human readable license, but also

provides a machine readable version of the license, enabling computers to determine in what way the content may be reused (Lessig, 2004). Machine readable licenses enable others to create automated services, based on the type of reuse granted by the content owner.

Although both Creative Commons and the open access movement seek to restore the balance between the owners of creative works and the prospective users, not all Creative Commons licenses are considered equally 'open' by OA proponents. For example, the BOAI now recommends a specific CC license: CC-BY (Budapest Open Access Initiative, 2012). According to the open access Scholarly Publishers Association (OASPA) this license allows for unrestricted reuse of content, as long as the source work is appropriately attributed (Redhead, 2012).

The preference for this license is not undisputed[13]; and has led to discussions about the relative merits of the degree of openness provided by the different CC licenses. In this paper, we will use a simpler distinction: documents which are available without charge, and documents that are not only available without charge, but also made available under a license that enables reuse and further dissemination. Peter (Suber, 2008) coined the terms 'gratis' and 'libre' OA to distinguish between these two distinct forms of 'open access'. Throughout this paper, books published under *any* type of CC license are categorised as *libre* open access; all other freely accessible books are categorised as *gratis* OA. In other words, books published under open licences ranging from CC-BY to CC-BY-NC-ND are here defined as *libre*;[14] books which are only 'free to read' and may not be freely used otherwise are defined as *gratis*.

Because documents which have been released under a *libre* license – such as Creative Commons – carry fewer barriers to reuse, it stands to reason that such content is easier to share and more likely to be used. This paper examines a discrete collection of open access monographs – the OAPEN Library collection – in an attempt to determine whether *libre* licenses do in fact lead to greater use of open access works.

13 A recent example is the interview with Paul Royster (Poynder, 2014).

14 Although the most restrictive CC licenses do not permit the adaptation of content, they still allow greater reuse than a *gratis* license that restricts users to the personal use rights under copyright law.

5.2 The OAPEN Library and the DOAB

The OAPEN Library was officially launched in September 2010 (OAPEN Consortium, 2011). It is a web-based collection of open access monographs, published by dozens of publishers. In December 2013, the collection contained over 2,000 titles by 55 publishers. The OAPEN Library offers several ways to make its contents accessible: it enables searching and browsing, readers can share book descriptions via social media, and it also offers several data feeds (Open Access Publishing in European Networks, 2010b). In the OAPEN Library, books are made available under several licenses: roughly 50% of the titles are disseminated under a Creative Commons license, while the rest are made available under a more restrictive regime. In other words, about half the titles in the OAPEN Library are available as *gratis* OA, the other half as *libre* OA.

The OAPEN Library is managed by the OAPEN Foundation. In April 2012, the OAPEN Foundation launched the Directory of Open Access Books (DOAB) as a discovery service for open access books ("DOAB: Directory of Open Access Books," n.d.). The directory is open to all academic publishers and aims to contain as many books as possible, provided that these books are peer-reviewed and published under an open access license. In addition to the publishers already taking part in OAPEN, several other academic publishers have placed their books in the DOAB (Snijder, 2013a). The DOAB is connected to the OAPEN Library, and automatically uploads descriptions of new titles from OAPEN. However, not all books in the OAPEN Library are listed in the DOAB: it only contains the titles with a Creative Commons – or a comparable – license. The selection is not limited to CC-BY, but extended to the full range of CC licenses. So, while the OAPEN Library is a mixture of *gratis* and *libre* OA, the DOAB only lists *libre* OA books. This automated aggregation – based on the machine-readable licence information – results in extra exposure of the *libre* books in the OAPEN Library.

5.3 Examining the Impact of Licenses on use

The OAPEN Library and the DOAB are examples of two types of use of open access works: use by individual end users and use by intermediaries, who provide additional services built on or around open content. Here, the end users are the readers of the books contained in the OAPEN Library. Use by this group can be measured by counting the number of times a book has been downloaded from the OAPEN Library. Downloads have been

chosen as a metric for two reasons. First, readers both within and beyond universities are able to download books held in the OAPEN Library. This contrasts to approaches to measuring impact that are based on counting citations, which only capture a specific form of use by academic readers. Second, downloads can be measured directly from the OAPEN server. This ensures a fast, and dependable, result. Although it is not possible to equate a download with further use (e.g. reading, integration into other work), we can assume that a much-downloaded online monograph has been read more often than a book which has been downloaded just a few times. We cannot, however, state that 100 downloads equate to 100 people reading the book cover to cover.

While all books in the OAPEN Library are openly available to download by end users, a significant proportion are also available under *libre* open access licenses. These types of licenses allow intermediaries like the DOAB to aggregate books and display them on a website, which creates another access point for individual users. This type of aggregation would not be possible with books available under a more restrictive *gratis* license.

In this paper, we examine both types of use – by individual end users and by intermediaries – as we consider the effects of *gratis* and *libre* licenses on the number of times books in the OAPEN Library are downloaded. In looking at this, we will make two comparisons: first, between *gratis* and *libre* books that were available in the OAPEN Library prior to the creation of the Directory of Open Access Books; and, second, between *gratis* books only available in the OAPEN Library and *libre* books available in the OAPEN Library and also included in the DOAB. This will allow us to measure whether *libre* books are downloaded more often in general, as well as whether additional aggregation has a significant effect on downloads. Our guiding research question for this study is:

Compared to *gratis* access, does applying an open license (*libre* access) have a positive effect on the number of times an open access book is downloaded?

Although the question of license effect is of primary interest, we are aware that the use of books in the OAPEN Library may also be influenced by factors other than license type, such as the subject or the language of the monographs. Earlier research published by (Snijder, 2013b) described the differences in number of downloads per subject in 2011. It seems reasonable to assume that subject still affects downloads. We could also argue that the language of a publication acts as a barrier to use: when readers cannot understand the language, the books become useless to them. And because

of the length of the texts involved, the chances of successful automatic translation are slim.

The number of times a book is downloaded might reasonably be expected to reflect the size of a particular language community. Therefore, it is important to remain alert to the impact of both subject and language on use when attempting to understand this ecosystem. Regardless of the type of license applied to a work, and whether or not it is made available in an aggregation service like the DOAB, prospective readers are not very likely to download books on subjects that are of no interest to them, or written in languages they cannot read.

Finally, it may be useful to note that information about the license for each individual book is always available to the users of the OAPEN Library website: each page describing a monograph contains a description of the license. Moreover, information about the complete collection can be obtained through several data feeds. On top of a description of the books, all data feeds also list the license information. Within the OAPEN Library, there is no technical distinction between books with a *libre* license or a more restrictive license: each monograph can be searched and downloaded in the same manner. Differences in usage cannot, therefore, be accounted for by restrictions in the infrastructure.

5.4 Literature review

There are three areas of literature relevant to this study: the conflicting interests of creators and users; the use of Creative Commons licenses to balance these interests, and the impact of Creative Commons licenses on usage.

5.4.1 Tensions between the interests of creators and users

Much of the debate around intellectual property, particularly copyright, centres on the tensions between creators' rights and users' rights. A much cited example is the paper by (Landes & Posner, 1989), in which they discuss the optimal level of copyright protection. This entails balancing the interests of the creators of a work versus the people who want to use it – either as a 'reader' or for creating a derivative work. The conflicting interest of these stakeholders is also described by (Boldrin & Levine, 2002). In their analysis, intellectual property law has two components. The first component is the right to own ideas and sell them. The second component

is the right to control the use of those ideas after sale. They argue that the second component – termed "downstream licensing" – leads to monopolies, impairing economic welfare. Again, we see the need to balance the interests.

Rather than finding a balance in current copyright law, (Suzor, 2014) argues that in certain cases, a high level of copyright protection is not needed. According to Suzor, many content users are prepared to pay the producers, even if the content is freely available. Choosing an intellectual property model that allows free use, while encouraging – but not enforcing – financial support may both enhance dissemination and compensate producers.

5.4.2 Balancing interests using Creative Commons licenses

Several authors have discussed the legal context around Creative Commons Licenses. Loren (2007) criticizes the "climate of overly broad ownership rights for creative works", and argues that it hinders the use and reuse of creative works. The complexity of the current copyright system leads to high costs, which disadvantages individuals who do not have the same financial resources as corporations. Broadly applying Creative Commons licenses helps to create a "semicommons of creative works" (Loren, 2007, pag. 328), which enables a greater and more diverse usage – to the benefit of society. This argument closely resembles the removal of legal barriers in the Berlin Declaration and the Bethesda Statement, describing a right to access and reuse scholarly and scientific content.

Hietanen (2008) also describes the advent of Creative Commons licenses as a reaction to the way copyright law has developed. Hietanen discusses the implications of applying CC licenses in great detail and analyses the license-choosing process and the clauses of the Creative Commons licenses. The approach by Kim (2007) is slightly different, and tries to understand the motives of CC licensors through surveys and interviews. Again, the conflict of interests of the different stakeholders are debated. However, Kim's paper attempts to categorize the types of content licensed under Creative Commons, and the motives of the content owners. The paper describes a large variety of content types. Furthermore, the reasons to use a CC license vary: some content owners place emphasis on the public benefits, while others are motivated by more personal reasons. (Morrison, 2012) discusses the application of CC licenses within open access publishing. According to the author, the goals of OA publishing and CC licenses are not aligned. She concludes that the lack of restrictions of the CC-BY license actually might be harmful to OA; the absence of restrictions leaves the author or content

owner without tools to control its reuse – suggesting that some licenses may tip the balance too far.

5.4.3 Do Creative Commons licenses enhance usage?

Despite Morrison's (2012) concerns, other authors arrive at more optimistic conclusions regarding Creative Commons licenses. (Carroll, 2006) looks at CC licenses and the changing role of intermediaries. The licenses are made machine readable, which opens new possibilities for those who enable all kinds of transactions based on the licensed works. The image sharing website Flickr.com is a well-known example: it enables end users to find pictures published under licenses that allow reuse. (Guibault, 2011) discusses the relation between authors of scientific and scholarly works and copyright ownership in the European context. She concludes that licensing documents under Creative Commons (partly) increases access and reuse.

There is little to no research published on the effects of *gratis* versus *libre* open access, especially in the realm of monographs. (J. L. Hilton III, Lutz, & Wiley, 2012) investigated revisions made to academic textbooks published under an open license. They conclude that – in line with expectations – the amount of revisions is relatively low. This is consistent with the findings in this paper; an open license does not automatically lead to a surge in usage. As far as could be established, there is no literature available which aims to quantify variances in usage based on differences in licenses.

5.5 Methods and the data set

The OAPEN Library platform logs usage data, starting from January 2011. Among the data recorded is the number of times each monograph has been downloaded in a month. We will use this as an indicator of successful dissemination: more downloads means a better result. For this paper, we will analyse the data captured over a period of 33 months: from January 2011 up until September 2013. During this time, 1,734 different books were made available through the OAPEN Library. Of these monographs, 855 were disseminated as *libre* open access and 879 were distributed under a more restrictive regime. Of the 855 *libre* titles, 512 were published under a CC-BY-NC-ND licence; the most restrictive open license. In contrast, only 4 titles were available under the CC-BY license. The rest of the titles were licensed as follows: 162 titles under CC-BY-ND; 168 titles under CC-BY-NC; and 9 titles under CC-BY-SA.

The Directory of Open Access Books was launched in April 2012, 16 months after January 2011. To understand whether the DOAB influences usage, we will compare the data of the first 15 months of the OAPEN Library to the data of the following 18 months. In the first 15 months, a total of 935 monographs were disseminated via the OAPEN Library; 563 of those under a *libre* license. After that period, the collection grew to the 1,734 books, as described above. The monthly download data for each book is used; if a book has been available for 33 months, this leads to 33 samples. Of course, not all books were available during that period, but the total number of samples used in our analysis is over 34,000.

Table 1 lists the number of books that were made available in the OAPEN Library, split by period and license. In the period before the launch of the Directory of Open Access Books, the difference in usage is not very large: on average, books published under a *libre* license were downloaded 29 times per month, compared to 21 times per month for books with *gratis* licenses. However, in the period after the DOAB launch, the difference widens to 84 downloads versus 34 downloads on average per month. It seems reasonable to assume that the aggregation in the DOAB has a positive influence.

Table 1 Average downloads per period

Period	Libre OA		Gratis OA	
	Number of books	**Mean downloads (SD)**	**Number of books**	**Mean downloads (SD)**
Direct use only (Jan. 2011-Mar. 2012)	563	29.6 (66.0)	372	21.9 (37.4)
Aggregation and/or direct use (Apr. 2012-Sep. 2013)	855	84.1 (409.1)	879	34.5 (44.7)

The OAPEN Library contains books on many subjects; our dataset contains 96 different subject classifications. Nevertheless, not all subjects are equally spread among the collection. Among the most common subjects we find Politics & Government and History. When looking at the licenses used, it becomes clear that they are not evenly spread: for instance, 22% of the books on Politics & government are published under a *libre* license, compared to 61% of books on History. Table 2 contains a more comprehensive listing.

Table 2 Subjects in the OAPEN Library

Subject (BIC classification)	Total number of books	Percentage	Books: *libre* license	Percentage (of all books)	Books: *gratis* license	Percentage (of all books)
Politics & government (JP)	398	23.0%	93	5.4%	305	17.6%
History (HB)	237	13.7%	144	8.3%	93	5.4%
Society & culture: general (JF)	129	7.4%	75	4.3%	54	3.1%
Economics (KC)	107	6.2%	14	0.8%	93	5.4%
Sociology & anthropology (JH)	77	4.4%	24	1.4%	53	3.1%
Other subjects	786	45.3%	505	29.1%	281	16.2%
Total	1,734	100%	855	49.3%	879	50.7%

The collection contains monographs in several languages. Most are written in English, Dutch, German or Italian, but also books in Danish, Latin or Russian are made available. As is the case with subject, the portion of books published under a *libre* license varies strongly per language; while 57% of books in English can be downloaded using a *libre* license, the percentage for Dutch is much lower: 13%. Table 3 lists the number of books per language.

Table 3 Languages in the OAPEN Library

Language	Total number of books	Percentage	Books: libre license	Percentage (of all books)	Books: gratis license	Percentage (of all books)
English	711	41.0%	408	23.5%	303	17.5%
Dutch	494	28.5%	62	3.6%	432	24.9%
German	346	20.0%	303	17.5%	43	2.5%
Italian	118	6.8%	74	4.3%	44	2.5%
Other languages	65	3.7%	8	0.5%	57	3.3%
Total	1,734	100%	855	49.3%	879	50.7%

The complete data used for this paper is available at http://persistent-identifier.nl/?identifier=urn:nbn:nl:ui:13-8ut1-25.

5.6 Analysis

Our analysis starts with measuring the effect of four factors – license, DOAB aggregation, subject, and language – on usage. This helps to determine if all factors indeed affect the number of downloads in the OAPEN Library. If one or more of them is not relevant, it can be discarded from our analysis. The one-way independent ANOVA statistical method is used to check whether each influence has a statistically significant effect. This procedure tests if the differences between the mean downloads of the books can be explained by chance. The results of each individual test are summarized in Table 4.

Table 4 Effects of the factors

Influence	Results
License	There was a significant effect of license on monograph downloads, $F(1, 19575.517) = 195.114$, $p < .001$, $\omega^2 = 0.00$
DOAB aggregation	There was a significant effect of DOAB aggregation on monograph downloads, $F(1, 25226.413) = 277.956$, $p < .001$, $\omega^2 = 0.00$
Subject	There was a significant effect of subject on monograph downloads, $F(10, 5995.946) = 46.935$, $p < .001$, $\omega^2 = 0.00$
Language	There was a significant effect of language on monograph downloads, $F(4, 10528.836) = 248.871$, $p < .001$, $\omega^2 = 0.01$

Results: The assumption of homogeneity of variance has been violated; therefore, the Welch F-ratio is reported.

When a statistically significant effect has been measured, the differences between the analysed groups is bigger than can be expected by chance. However, in large groups this will happen more often; our data set contains over 34,000 samples. The height of the *F*-ratio indicates the effect size: a higher ratio indicates a stronger effect of the experiment – in our case: license; aggregation through the DOAB; subject and language. Furthermore, the ω^2 value describes the proportion of the variance between the two groups. If the value of ω^2 is 0.01, this means that approximately 1% of the difference in downloads can be attributed to the effect investigated.

The results show that usage of the OAPEN Library is not only influenced by license; it is also affected by DOAB aggregation, subject and language. This complicates the goal of identifying the specific influence of license type. A common way to proceed is to use the multifactor ANOVA procedure to measure the effect of license, combined with the impact of DOAB aggregation, subject and language. Nevertheless, in order to get meaningful results from this procedure, several requirements must be met. The most important precondition is the homogeneity of variance. In other words, the

means used in the procedure should be evenly distributed. Unfortunately, our data does not meet this condition. As a possible solution to overcome the statistical problems, the data is split into smaller subsets.

We have seen before that the usage in the period before the DOAB launch is strongly different from the usage patterns in the period after the launch. To compensate for this, the data is split in two sets: the usage statistics generated in the period before the launch of the directory – January 2011-March 2012 – and the number of downloads registered in the period when DOAB aggregation was deployed – April 2012-September 2013. The groups to be analysed share the same subject or the same language, and the data was gathered in the same period. So, for instance, the usage of all books with the subject Politics & government in the period before the launch of the DOAB – January 2011-March 2012 – is analysed to see whether the license has a significant influence. Splitting up the data creates smaller subsets; but even the smallest group – Sociology & anthropology, in the period January 2011-March 2012 – contains 415 samples.

The analysis focuses on the impact of licensing on direct use (usage pre-DOAB launch), and on the impact of licensing on direct use and aggregated use (usage post-DOAB launch). When we look at use prior to the launch of the DOAB, we expect that simply using *libre* licenses will have a positive effect on the number of times books are downloaded by readers. When we examine use after the launch of the Directory of Open Access Books, the role of a *libre*-enabled intermediary in providing an additional access point is analysed. With regard to aggregators like the DOAB, it is expected that *libre* licenses will enhance the number of downloaded books indirectly--by facilitating additional access points which stimulate readers to find and download books.

As we have seen, the results are not only affected by the license used; the effects of subject and language also play a prominent role. The effects of language and subject are not straightforward: whether or not a certain language or subject enhances or diminishes the number of books downloaded is hard to predict. In contrast, the use of a *libre* license is directly aimed at removing barriers to usage. The impact of subject and language can be seen in the analysis below in that the influence of licenses varies per dataset. However, the overall picture is clear: the use of *libre* licenses alone has a limited impact on downloads, while aggregating *libre*-licensed books has a positive effect on the number of books downloaded.

5.6.1 Impact of licensing on OAPEN downloads

5.6.1.1 Subjects and license

Here, the difference in mean number of downloaded books is examined between *libre* and *gratis* books that share the same subject. This analysis only includes usage prior to the launch of the DOAB. The results are mixed: for the books on History or the books on Society & culture, the license has no effect on the number of books downloaded. However, for the other subsets, the differences in mean number of downloaded books is statistically significant.

Even where there is a significant difference, the effects of publishing under an open license are not very large. Before, we discussed the *F*-ratio and the ω^2 value as an indication of the impact. If we look at these numbers, it becomes clear that the effect of *libre* licenses for books on Economics is much smaller compared to the other subsets. Also, the ω^2 value is never higher than 0.02. In other words: *libre* licenses do not always lead to a difference on the number of books downloaded; when such a difference is found, the influence of licences is much smaller for books on Economics and for other groups the measured impact is no more than approximately 2%.

Table 5 lists the mean number of downloads per subject in the time before the launch of the DOAB.

Table 5 Subjects and license; direct use only

Subject	*Libre* license		*Gratis* license		Results
	N	Mean downl. (*SD*)	N	Mean downl. (*SD*)	
Politics & government (JP)	969	26.6 (36.8)	1169	17.3 (22.2)	There was a significant effect of license on monograph downloads, $F(1, 1525.148) = 47.376$, $p < .001$, $\omega 2 = 0.02$
History (HB)	1136	20.6 (22.3)	785	21.5 (28.9)	No significant effect of license on monograph downloads could be found, $F(1, 1919) = 0.7$, $p = .403$, $\omega 2 = 0.00$
Society & culture: general (JF)	635	46.8 (171.1)	263	37.6 (107.4)	No significant effect of license on monograph downloads could be found, $F(1, 896) = 0.655$, $p = .418$, $\omega 2 = 0.00$

Subject	*Libre* license		*Gratis* license		Results
	N	Mean downl. (*SD*)	N	Mean downl. (*SD*)	
Economics (KC) 213		41.4 (81.1)	569	25.8 (31.7)	There was a significant effect of license on monograph downloads, $F(1, 236.640) = 7.446$, $p = .007$, $\omega 2 = 0.00$
Sociology & anthropology (JH)	356	30.1 (34.8)	59	14.5 (13.1)	There was a significant effect of license on monograph downloads, $F(1, 221.856) = 38.333$, $p < .001$, $\omega 2 = 0.00$
Other subjects	3466	29.5 (44.7)	2116	21.7 (30.8)	There was a significant effect of license on monograph downloads, $F(1, 5502.144) = 58.887$, $p < .001$, $\omega 2 = 0.01$

Results: With the exception of "History (HB)" and "Society & culture: general (JF)", the assumption of homogeneity of variance has been violated; therefore, the Welch *F*-ratio is reported.

5.6.1.2 Languages and license

Here we follow the same procedure: the data is split into groups with the same language in order to create groups with equal attributes. The data in Table 6 was captured before launching the DOAB.

Again we see that license type does not create a statistically significant difference in all groups, and that both the *F*-ratio and the ω^2 value are relatively low in the groups where a statistically significant difference is found. The maximum ω^2 value is even lower compared to the analysis on subject: it is 0.01. In other words, the biggest measured impact of licenses is approximately 1%. Moreover, the books written in Italian and other languages – where no significant statistical differences were found – show a different download pattern: the mean downloads of books with a *libre* license is lower compared to the group of *gratis* titles.

Table 6 Languages and license; direct use only

Language	Libre license		Gratis license		Results
	N	**Mean downl.** (SD)	**N**	**Mean downl.** (SD)	
English	3883	35.4 (83.3)	2233	27.6 (47.9)	There was a significant effect of license on monograph downloads, $F(1, 6113.989) = 21.867$, $p < .001$, $\omega^2 = 0.00$
Dutch	598	24.6 (24.9)	978	21.0 (29.0)	There was a significant effect of license on monograph downloads, $F(1, 1574) = 6.074$, $p = .014$, $\omega^2 = 0.00$
German	1221	26.5 (30.8)	433	20.2 (31.1)	There was a significant effect of license on monograph downloads, $F(1, 752.804) = 13.153$, $p < .001$, $\omega^2 = 0.01$
Italian	1052	14.9 (25.1)	586	16.3 (21.7)	No significant effect of license on monograph downloads could be found, $F(1, 1357.292) = 1.382$, $p = .240$, $\omega^2 = 0.00$
Other languages	21	9.0 (9.8)	731	10.9 (12.5)	No significant effect of license on monograph downloads could be found, $F(1, 750) = .492$, $p = .483$, $\omega^2 = 0.00$

Results: With the exception of "Dutch" and "Other languages", the assumption of homogeneity of variance has been violated; therefore, the Welch F-ratio is reported.

5.6.1.3 Conclusion on the impact of licenses on downloads

From a statistical point of view, the number of downloaded books is sometimes positively affected by open licenses. However, we have also seen that if there is a positive effect, it is very small. Furthermore, not all groups of books are affected by the license. If the books are grouped by subject, for the titles on History and the books Society & culture – 21% of all titles – the difference in number of books downloaded is not caused by the license. When the books are grouped by language, we see a statistically significant effect for monographs written in English and Dutch – almost 70% of all titles – with an associated ω^2 value of 0.00. An effect of approximately zero percent is not very large.

We can conclude that the impact of *libre* licenses is limited – the download behaviour of users of the OAPEN Library is not affected in any practical way by the type of license used. However, in the next section we will see that *libre*-enabled aggregation through an intermediary has a much bigger effect on usage.

5.6.2 Impact of license-enabled aggregation on OAPEN Downloads

5.6.2.1 Subjects and aggregation

When we look at the download data for the period after the launch of the DOAB, the results are quite different. Compared to their *gratis* counterparts, each group of monographs published under a *libre* license and so listed in the DOAB is downloaded more. Here, the mean number of downloads of books under a *libre* license is almost twice as high compared to *gratis* titles. In the previous data set, the difference is closer to 25%.

In addition, not only are the differences in mean downloads larger, but the statistical effects are also more profound. First, the *F*-ratios – defining the size of the effect we are measuring – are much higher compared to the data set listed in Table 5. Also, the values of ω^2 are much bigger. In the case of Sociology & anthropology it is 0.17; about 17% of the difference could be explained by the *libre* license and the subsequent aggregation through the Directory of Open Access Books. Table 7 lists the data of the monographs grouped by subject.

Table 7 Subjects and license; aggregation and direct use

Subject	Libre license (Access: OAPEN and DOAB)		Gratis license (Access: OAPEN only)		Results
	N	Mean downl. (SD)	N	Mean downl. (SD)	
Politics & government (JP)	1812	69.6 (54.4)	1516	34.8 (37.7)	There was a significant effect of license on monograph downloads, $F(1, 3218.685) = 468.751$, $p < .001$, $\omega^2 = 0.12$
History (HB)	1507	88.3 (159.6)	2894	24.3 (28.0)	There was a significant effect of license on monograph downloads, $F(1, 1554.432) = 237.930$, $p < .001$, $\omega^2 = 0.09$

Subject	Libre license (Access: OAPEN and DOAB)		Gratis license (Access: OAPEN only)		Results
	N	Mean downl. (SD)	N	Mean downl. (SD)	
Society & culture: general (JF)	1109	87.4 (64.4)	666	42.8 (42.6)	There was a significant effect of license on monograph downloads, $F(1, 1756.454) = 306.974$, $p < .001$, $\omega^2 = 0.12$
Economics (KC)	352	99.0 (62.1)	849	39.6 (38.4)	There was a significant effect of license on monograph downloads, $F(1, 466.097) = 276.973$, $p < .001$, $\omega^2 = 0.10$
Sociology & anthropology (JH)	757	73.7 (55.9)	72	35.1 (24.7)	There was a significant effect of license on monograph downloads, $F(1, 153.424) = 117.562$, $p < .001$, $\omega^2 = 0.17$
Other subjects	6491	86.9 (549.6)	4386	38.8 (55.2)	There was a significant effect of license on monograph downloads, $F(1, 6683.120) = 49.062$, $p < .001$, $\omega^2 = 0.00$

Results: The assumption of homogeneity of variance has been violated; therefore, the Welch F-ratio is reported.

Languages and aggregation

When the titles are grouped by language, the statistical effects of a *libre* license leading to aggregation by the DOAB are also visible. Most interesting are the differences in *F*-ratios and ω^2 values between the different language groups. While the *libre* titles written in Dutch and the titles written in "Other languages" clearly benefit from the aggregation, the effects on books in English and German are less noticeable. Still, the findings are statistically significant, and another metric is also clearly pointing in the same direction. If only direct usage is analysed – the data in Table 6 – the difference between mean number of downloads of books on a *gratis* licence is small; the average amount of downloaded *gratis* books is almost as high as the mean number of downloads of books on a *libre* license. However, the data in Table 8 depicts a much larger difference. Here, the mean number of downloads of *libre* books is almost twice the amount for *gratis* books.

Table 8 Languages and license; aggregation and direct use

Language	Libre license (Access: OAPEN and DOAB)		Gratis license (Access: OAPEN only)		Results
	N	Mean downl. (SD)	N	Mean downl. (SD)	
English	6245	118.3 (565.4)	4018	50.7 (51.7)	There was a significant effect of license on monograph downloads, $F(1, 6406.638) = 88.388$, $p < .001$, $\omega^2 = 0.01$
Dutch	962	55.8 (35.5)	4031	22.9 (38.5)	There was a significant effect of license on monograph downloads, $F(1, 1547.706) = 644.752$, $p < .001$, $\omega^2 = 0.10$
German	3466	47.7 (39.3)	674	36.6 (42.7)	There was a significant effect of license on monograph downloads, $F(1, 907.778) = 38.820$, $p < .001$, $\omega^2 = 0.01$
Italian	1258	37.1 (40.1)	748	21.6 (24.9)	There was a significant effect of license on monograph downloads, $F(1, 2000.270) = 113.112$, $p < .001$, $\omega^2 = 0.04$
Other languages	97	63.5 (51.4)	912	22.9 (22.1)	There was a significant effect of license on monograph downloads, $F(1, 99.828) = 59.341$, $p < .001$, $\omega^2 = 0.14$

Results: The assumption of homogeneity of variance has been violated; therefore, the Welch F-ratio is reported.

5.6.2.2 Conclusions on the impact of license-enabled aggregation on downloads

In contrast to the download activity prior to the launch of the Directory of Open Access Books, there is a statistically significant effect on all subsets: the use of an open licence, which allows the creation of an additional access point through the DOAB, has a positive effect on the number of books downloaded. The influence of aggregation clearly makes a difference. The most positive statistical effects are found within the subset "Sociology & anthropology" – where approximately 17% of the difference can be explained by open licensing and the subset "Politics & government" and the subset "Society & culture: general" – here approximately 12 % is measured.

However, not all results are so unambiguous, especially for the subsets on language. For instance, while a positive influence has been measured, the value of ω^2 for books in English is just 0.01. On the other hand, the mean

number of downloaded English language books on a *gratis* license is less than half the mean number of books on a *libre* license.

We can conclude that the use of *libre* licenses has a positive effect when we look at the effect of aggregation on downloads. Although the licenses do not directly affect the readers' behaviour, *libre* licences enable additional services by intermediaries like the Directory of Open Access Books. These additional services lead to increases in the number of books downloaded.

5.7 Discussion

The notion that *libre* material will be more used compared to *gratis* works seems highly obvious: an open license removes a barrier to usage. On the other hand, if the *gratis* works are made available under the same technical conditions as their *libre* counterparts, most users would make no distinction and treat the works as 'free as in beer'. In the case of the OAPEN Library, its description of licenses states the following: "If not stated otherwise, all works in the OAPEN Online Library fall under the OAPEN Deposit License – all rights reserved. End users are allowed to read the work online, download, print and copy it for their own personal purposes within the legal framework of their national copyright law. Beyond this all rights are reserved."[15] In other words, the site clears legal obstacles for readers who want to use the books for personal reasons, and in this context it is not surprising that *libre* licenses did not play a large role in the period before the launch of the DOAB (January 2011-March 2012).

We have seen that each of the four discussed influences – *libre* versus *gratis* licenses; additional aggregation; subject and language – all affect the usage of the books in the OAPEN Library. By looking at the period before the extra coverage provided by the Directory of Open Access Books could play a role, a possible influence is removed from the analysis. As a second restriction, the usage data is split among subjects or languages. Within some of these subsets, the *libre* license positively affects usage, while in other subsets the effect could not be measured. However, even if a statistically significant result has been found, the effect size was negligible. The biggest measured impact of licenses found in the analysis of the subject subsets is approximately 2%. If languages are examined, almost 70% of all titles listed an effect of approximately zero percent. These results refute the claim by Guibault (2011) that open licenses enhance usage. However, in this

15 http://oapen.org/about?page=support&subpage=forreaders

particular case, the legal restrictions toward books with a more restrictive license are relatively slight.

Combining *libre* licenses and aggregation in the DOAB has a far more profound effect. When the data of that period is split in subsets based on subject or language the difference is clear. In each subset, the books with a *libre* license are downloaded more; the additional access provided through the DOAB appears to result in more successful dissemination of the books. This is also seen in the ratio between the mean number of downloads before and after the deployment of the DOAB. Taking into account all the average downloads in the subject subset reveals that in the pre-DOAB period, the number of downloads for books with a *gratis* license is 72% of the amount associated with books published under a *libre* license. After the launch of the DOAB, this percentage plummets to 43%. The same holds true in the language subset, where the percentages are 91% and 54%, respectively. This is another indication that extra aggregation has a positive impact on usage.

5.8 Conclusion

As far as could be established, this is the first paper to measure the effects of *libre* licenses on the use of open access monographs. Most of the literature on open licenses discuss them from a legal perspective, and focus on their innovations in relation to copyright. Also, open access publishing as a means to optimize the dissemination of scholarly and scientific information is mostly absent from the articles cited. However, the underlying theme – ownership and control over creative works and its economic aspects – does of course play an important role in the OA debate. Enforcing restrictions based on copyright laws creates another barrier to access, or to certain types of reuse.

Both the open access movement and the Creative Commons organization strive to maximise the use of creative works. While they share the goal of removing legal barriers to use or reuse, there is disagreement about the optimal license for open content. The Creative Commons organization chooses a flexible approach, by offering six different choices. In contrast, within the open access movement, there is a strong preference for the CC-BY license.

The current collection of the OAPEN Library does not completely conform to the recommendations of either group. Roughly half of the collection is made available under a *gratis* license that only permits personal use, which is more limited than the most restricted Creative Commons

license. Nevertheless, when considering direct use only (pre-DOAB launch downloads), the books under a *gratis* license perform just as well as the *libre* titles. In this context, the impact of licenses is limited.

However, when examining the use of OAPEN Library books after the launch of the DOAB, which automatically imports metadata of all books with a *libre* license, a benefit of *libre* licenses becomes clear. As Carroll (2006) predicted, machine readable metadata on licenses was used to perform a service; in this case inserting the OAPEN titles into the DOAB discovery service. Doing so proved to be successful: the titles featured in the DOAB are downloaded from the OAPEN Library more compared to books which do not receive the extra attention.

To a certain extent, the decision to include *libre*-licensed OAPEN titles in the Directory of Open Access Books – leading to additional visibility on another platform – has been a DOAB policy decision, and was not inherently dependent on license type. However, the machine readable *libre* licenses that enable aggregators such as the DOAB to identify and add licensed content can also lead to other types of reuse. For example, Biomed Central offers text mining services based on a collection of articles with a "BioMed Central open access license agreement". According to BioMed Central, this license is identical to the Creative Commons Attribution License (BioMed Central Ltd., 2014).

Whether through simple aggregation or more intensive reuse like textual analysis, it appears that *libre* licenses do have the potential to positively affect usage. Rather than directly appealing to end users of individual books, these licenses enable intermediaries to create new services built on collections of open content. These services, in turn, can help to increase the impact of the individual publications.

5.9 Limitations

In the data set used for this paper, each book's license was described in two ways: Creative Commons or no Creative Commons. It did not take into account the six different licenses in several versions – 2.0, 2.5 and 3.0 – that have been used in the examined collection. Some of the books were published under the UK or German version, while most were published under the 'international' version. It may be possible that the readers of the OAPEN Library were aware of all the legal details, and this influence has not been taken into account. The metadata of the books – available at http://

persistent-identifier.nl/?identifier=urn:nbn:nl:ui:13-8ut1-25 – contains the license of each individual title.

In the statistical analysis, is has been assumed that the choice for publishing a book under a *gratis* or a *libre* licence has not been biased. The influence of license on the behaviour of readers has of course been extensively discussed.

5.10 Acknowledgements

The author would like to thank professor Paul Wouters of the Centre for Science and Technology Studies (CWTS) for commenting on the draft version of this paper and Marieke Polhout of Data Archiving and Networked Services (DANS) for making the data available. Furthermore, the support of the publisher and the copy-editor have been crucial for the publication of this paper.

6 Patterns of information : Clustering books and readers in open access libraries

6.1 Introduction

Open access libraries operate in a continuum between two distinct organisation models: online retailers versus 'traditional' libraries. Online retailers such as Amazon.com are successful in recommending additional items that match the specific needs of their customers. The success rate of the recommendation depends on knowledge of the individual customer: more knowledge about persons leads to better suggestions. Thus, to optimally profit from the retailers' offerings, the client must be prepared to share personal information, leading to the question of privacy.

In contrast, protection of privacy is a core value for libraries. The question is how open access libraries can offer comparable services while retaining the readers' privacy. A possible solution can be found in analysing the preferences of groups of like-minded people: communities. According to Lynch (2002), digital libraries are bad at identifying or predicting the communities that will use their collections. It is however our intention to explore the possibility to uncover sets of documents with a meaningful connection for groups of readers – the communities. The solution depends on examining patterns of usage, instead of storing information about individual readers.

This paper will investigate the possibility to uncover the preferences of user groups within an open access digital library using social networking analysis techniques.

6.2 Background

Recommender systems are powerful tools, whose design poses privacy issues. The role of privacy in the library landscape is discussed, along with the use of recommendation systems in libraries. If it is not feasible to match titles to individuals, the use of clustering techniques might mitigate some

of the privacy problems while still creating relevant sets of titles. In turn, these sets may be used in recommendation services.

6.2.1 Recommender systems

Recommender systems can be defined as tools that provide suggestions about items that may prove valuable to a user (Linden, Smith, & York, 2003; Pazzani & Billsus, 2007; Ricci, Rokach, Shapira, & Kantor, 2011; Schafer, Konstan, & Riedl, 1999). The prediction is based on processing data about items, users and transactions. Items are the objects to be recommended; in the case of digital libraries this would be documents. Understanding users is a critical part of recommender systems; ultimately, their success is based on how well they know the user's needs and preferences. Needless to say, this poses privacy issues. Transactions are defined as a recorded interaction between a user and the recommender system.

Recommender systems are based on several techniques. The first type of system is content based, in which recommendations are based on items that are similar to those used in the past. Another type of recommender system is based on the demographic profile of the user. A third kind deploys specific domain knowledge: what aspects of items are the most useful in a particular environment? Community based systems use recommendations of the user's friends. Finally, hybrid systems combine several of the discussed techniques. The common factor is creating an extensive profile of users at the level of the individual not limited to their personal preferences, and including data about their peers. Furthermore, this profile is updated over time to keep abreast of changing preferences. From a privacy point of view, this leads to the question of trust: how much personal information should such a system contain?

Trust in recommender systems has been investigated by Chellappa & Sin (2005), from a slightly different angle: under what conditions are people willing to allow vendors to store personal information? They conclude that people are prepared to share information if the vendor is able to invoke trust. The level of trust invoked by a specific vendor is a reason for consumers to shop there, and ignore others with virtually the same offering. Even while people feel a general concern about sharing private data in general, they might be willing to give up some of their privacy in return for benefits provided by the vendor.

Not everybody will be trading privacy for convenience, and Jeckmans *et al.* (2013) have investigated possible remedies, such as raising awareness about privacy issues and invoking specific laws dealing with personal

information. These types of measures have serious drawbacks. As we have seen, being aware does not stop people to engage with recommender systems and most legislation will take quite some time before coming into effect. The authors also describe technical measures such as anonymization, randomization and the use of cryptography. If user data is anonymized, the identifying information is removed, while preserving the rest. Randomization and differential privacy techniques aim to make the data of a specific person indistinguishable from most other users, by adding random data. Cryptography is considered to be a more secure choice, but with additional costs: it requires extra resources and may slow down the system.

These techniques add extra complexity to the system. This raises a question for the system's owner that mirrors the privacy trade-off by customers. Improved privacy protection will most likely have a negative effect on the system's efficiency, reducing the likelihood of implementation.

6.2.2 Libraries, privacy and the role of the catalogue

Global library cooperative OCLC lists at the time of writing 139 web based collections of open access documents (OCLC, 2016). All these collections fall within the definitions of digital libraries as discussed by Borgman (1999): a combination of "content collected on behalf of user communities", which also functions as an "institution or service". So, when a digital library collects and maintains a collection of documents in order to serve the information needs of specific groups of users, it functions as a 'traditional' library.

If open access libraries share traits with traditional libraries, we might also expect the same attitude towards privacy. The privacy of library patrons must be protected, including user data collected in library systems. This position is shared among the International Federation of Library Associations and Institutions (2016), the American Library Association (2014) and several other national library associations..

Protecting library patron's privacy is not an easy task. American libraries struggle with the implication of the USA PATRIOT Act, which expand the abilities of law enforcement agencies to collect personal information (Jaeger, McClure, Bertot, & Snead, 2004). The gathering of this type of data is not limited to the United States, but is also becoming more common in European countries (Nijboer, 2004). Apart from governmental organisations, libraries might also develop policies about other third parties who might be interested in the data generated by – and about – users (Corrado, 2007). Some libraries try to resolve the trade-off between extra functionalities and

better service versus protecting personal information by adding recommender functions to their online public access catalogue (OPAC), based on anonymised usage data (Geyer-Schulz, Neumann, & Thede, 2003; Mönnich & Spiering, 2008).

Whether a 'anonymized' OPAC is a fitting solution for open access libraries can be called into question. Firstly, the role of the catalogue as primary entrance to the collection is being re-evaluated, as illustrated by Dempsey (2006). He argues that library catalogues are too limited as tools to discover content. This is put into practice at the library of Utrecht University, through the deprecation of their OPAC system. Instead, relevant results must come from search engines and library aggregators (Kortekaas & Kramer, 2014). Others are discussing whether social media websites such as Facebook.com offer an alternative. Scale (2008) concludes that Facebook does not deliver optimal results, but the article's number of citations indicates the interest in the library community. Secondly, compared to 'traditional' libraries, open access libraries – which are by definition online – might even be more depending on search engines or other external discovery tools. This is illustrated by the OAPEN Library. Its website functions as an OPAC; however, over 70% of its usage bypasses the website. The documents are accessed by enabling integration into the user's systems, the infrastructure used on a daily basis (Snijder, 2014a).

In short, privacy should be a concern for open access libraries and the OPAC – even when it does not retain reader data – might not be the best solution for content discovery.

6.2.3 Clustering books and readers through social network analysis?

The previous sections made clear that recommendation systems only function well at the cost of privacy. In a library context, this is not acceptable and the offered solutions are not ideal, especially in the case of open access libraries. This leads to the central question to discuss in this paper: how to support library users in an environment that minimizes the amount of information stored about individuals?

When it is not feasible to create profiles of individuals, we might look at the combined behaviour of all users of the digital library. Are all books downloaded at random, or can we discern clusters of books that are meaningful for groups of readers? The clustered books can be seen as a network of interconnected objects. If it is possible to identify such networks, it might be possible to recommend relevant books based on usage patterns. We might go a step further, and examine if the groups of individuals connected to

those book clusters share a common trait. Thus, we are examining possible networks of books and readers.

How can we study such networks? Open access libraries do not register individual users, but a small amount of – publicly available – information about the internet provider can be used. Typically, the usage amounts to thousands of document downloads, where the provider and the document can be linked. In other words, the provider acts as a proxy for the reader. A certain document can be linked to multiple providers and one provider may be connected to multiple documents. These kinds of relations are studied using social network analysis techniques. Using graph – or network – theory, the characteristics of networks can be described and examined. Which aspects of the nodes – the parts – and the edges – the relations between the nodes – are most relevant depend on the characteristics of the network.

The possible combinations of providers and titles are quite large and thus, finding meaningful clusters is not easy. The same problem – at much larger scale – can also be found on the web. Kumar, *et al.* (1999) deployed graph theory to find "implicitly defined communities" using sets of interlinked Web pages. They aimed to find groups of content creators sharing a common interest. According to the authors, those groups could provide valuable information resources for interested users, uncover some of the sociology of the Web and target advertising. This aligns with the role of online libraries: providing valuable information to interested parties and directing them to the right documents is a core task. Finding communities in digital libraries is the first step to recommending useful content; not to individuals but to groups.

The extensive introduction into social network analysis by Wasserman & Faust (1994) can be used to define the type of network under examination. In this case, the network consists of two types of groups or modes: providers and documents. These types of networks are called two mode networks. Furthermore, the relation between the providers and the books is not reciprocated: providers act on books, but – for the purpose of this paper – books are not acting on the providers. Consequently, this two-mode network is directed.

Moreover, networks consisting of actors and passive elements such as social events – or in this case: documents – are called an affiliation or membership network. Here, the analysis is based on affiliations of actors to the passive elements, on the relation between the passive elements and the actors, or on both modes simultaneously. One possible analysis of the latter kind is finding cohesive subsets of actors and passive elements. In this case, clusters of providers and books.

The solution to the problem of finding communities in networks is described by Newman and Girvan (2004). By repeatedly using an algorithm that removes edges that acts as a 'bridge' between others, all the nodes are divided into closely connected groups. Wakita & Tsurumi (2007) have created an updated version, which is used for the research in this paper.

The use of social network analysis or clustering algorithms is not limited to the discovery of user groups. For instance, Verleysen & Weeren (2016) used a "fuzzy cluster analysis" to examine the divide between authors publishing in international journals in English, compared to those writing books and chapters in national or regional languages. Their results are supported by computer generated outputs. In contrast, Provan, *et al.* (2005) encourage community leaders to use social network analysis procedures to manually describe the networks they participate in. These approaches demonstrate the breadth of social network analysis.

The procedure outlined in this paper should be relevant to all kinds of digital open access libraries, leading to some additional requirements. First, it must be applicable to a wide variety of collections. Therefore, the metadata used has to be attainable from different types of documents. The metadata used will be discussed further in section *6.3.2 The books*. Secondly, the tools to be used ought to be available as open source software, preferably with an easy to use interface. This paper's analysis has been conducted using NodeXL, a free and open source network analysis tool using a Microsoft Excel template. It is maintained by the Social Media Research Foundation, co-founded by Marc Smith (Shneiderman & Dunne, 2013).

6.3 Quantifying the data set

The previous section discussed the tension between privacy and optimizing recommendation systems. Using social network analysis to find communities around certain books might enable open access libraries to create recommendations, while retaining the privacy of the individual readers. In order to test this idea, the usage of the OAPEN library will be analysed.

6.3.1 The collection

The OAPEN Library is managed by the OAPEN Foundation, a not-for-profit organisation based in the Netherlands. The Foundation's goal is to promote open access book publishing, through building and disseminating a quality-controlled collection of open access books (OAPEN Foundation,

2016). The books discuss a broad range of subjects, and are written in several languages. Around half the publications are written in English; both Dutch and German amount to roughly 20% and 5% of the books are Italian. Other languages include French, Danish, Spanish and Latin. Section *6.3.2 The books* describes the state of the collection in 2012.

As stated before, our goal is to find clusters of books and providers that have a meaningful connection. In other words, we need to establish whether a combination can be attributed to an underlying theme, and not determined by chance. This procedure must be transparent and reproducible in other collections than the one currently under examination. The method used is based on quantifying several aspects of the total collection. These numbers are compared to the amounts measured in the clusters.

The data describing the complete collection of 2012 and 2014 is available in the appendix. It will be used as benchmark. The data of the clusters described in section 6.4.1 and section 6.4.3 is also listed there. The appendix, the underlying lists of downloads, providers and the clusters they occupy are available via http://dx.doi.org/10.17026/dans-x72-d9h2.

6.3.2 The books

The clusters contain books and providers; the first step is to determine which aspects will be examined. Starting with books, the number of possible aspects is large. The books are collected and maintained by the owner of the digital library, who might choose to describe the documents in many different ways. A typical book description in a library catalogue contains the title, author, publisher, place and year of publication, number of pages, ISBN, language, whether it is part of a series, and indications of the book's subject through keywords and classification codes. However, these descriptions serve several purposes: some are useful to identify a work, while others may help to indicate the book's topic and its quality or prestige. In this case, we assume that the users of the electronic library are interested in books "about" a subject.

In general, the contents of a scholarly book will not be limited to one subject. Even if the authors are exploring one theme, the book will discuss several facets. An example is the book "Malaysian Cinema, Asian Film: Border Crossings and National Culture".[1] This book might be useful for those who are interested in film and media studies, but also for those who are involved in the culture of Malaysia or Southeast Asia.

1 See http://oapen.org/search?identifier=340243.

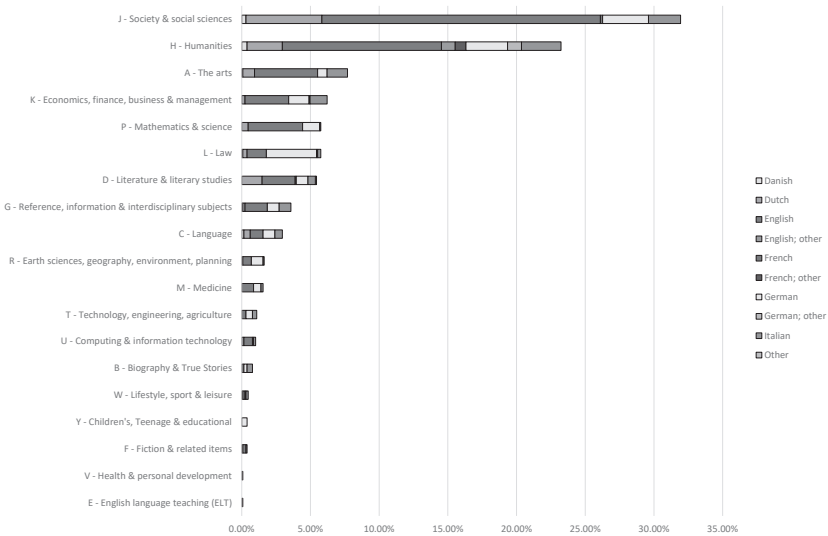
Another aspect to consider in a web-based collection is language. Open access libraries are open to everyone with an internet collection; a potential worldwide audience. However, it is unlikely that prospective readers will download books in languages they cannot understand.

The most widely used methods to describe the contents of document are keywords and classifications. Keywords are potentially more accurate, but the endless possibilities make it hard to create quantifiable sets. In contrast, classifications are based on a hierarchy. By selecting the top levels, it is possible to create relatively few sets of books.

Compared to keywords, the possible number of languages is lower. A large variation in languages is not always useful in an analysis. For instance, there is no point in listing Swedish as a separate language if a collection of thousands of titles contains three books in Swedish. Thus, the quantification is based on a simplification of the available metadata.

For the analysis, each book in the collection was categorised as follows: it belongs to one language set, and it may contain up to five subject codes. These two aspects can be placed in a matrix, serving as a 'snapshot' of the collection in a certain point of time. Such a matrix is quite useful in displaying every possible detail, and helps to quantify the amount of titles in a certain language or subject. Still, a visualisation is a more optimal way to display the main characteristics of the collection. For instance, the large percentage of documents on society and social sciences or humanities and the amount of English language books are easily spotted in *Figure 1*. Other details – such as the large percentage of German language books on law – are also visible.

Figure 1 OAPEN Library collection: languages and subjects (2012)



6.3.3 The providers

In contrast to books, the amount of information about providers is limited. When users are not tracked, there is little more available than the name of the provider and the time when a specific book was downloaded. Identifying individual readers is next to impossible: a feature that protects the reader's privacy. Using public information – based on the WHOIS internet protocol (Daigle, 2004) – the provider's country of origin is also available. Thus, using the provider as proxy, readers can be grouped by nation. This method is not 100% accurate: when a Dutch native travels through Canada and accesses a book using a Canadian internet provider, this will be listed as a "Canadian" download.

Logging internet providers leads to another interesting question: how many people have downloaded books through that provider, and are they interested in the same things? If the provider is an organisation with a strict focus, chances are that all members share a similar interest. An example is the organisation Bouwkennis – a Dutch marketing firm, specializing on the building sector, which downloaded a report on housing policy. In contrast, also listed among the downloading companies is Verizon, a large internet service provider serving millions of customers. It is highly unlikely that all

documents downloaded through Verizon are the result of a single person or a 'single minded' group. Another complication is the freedom of users of an online open access library to download as many titles as they like.

The question is how to select download patterns that are the result of a single person's action, or the action of a goal-oriented organisation. The number of titles downloaded by a provider should not be the deciding factor. On the contrary, we might imagine several readers who are interested in the same twelve books: the kind of pattern that hints at a shared interest. On the other hand, we need to filter out the actions of a diffuse group of people who only happen to share the same internet provider. The solution chosen here is to look at the number of times a single title is downloaded through a provider. The number of downloads are logged per month. To be absolutely sure that a single person has downloaded a title, multiple downloads of the same title by the same provider in one month are discarded. All downloads where just one copy of different titles is downloaded by a single provider are still part of the data. Besides, if the provider downloads the same title in another month, this download will also be part of the analysed data.

How does this choice effect the data? The download data for 2012 – collected during three months – consists of 6,176 providers who downloaded 57,508 books. After removing those providers that have downloaded the same title more than once in the same month, the number of providers becomes 5,180 (84% of 6,176) and the number of books downloaded is 34,345 (60% of 57,508). The majority (53%) of the 5,180 providers downloaded a single title; amounting to 2,740 providers. The remaining 47% (2,440 providers) downloaded between two and 338 different titles. Examining the number of providers that download more than one book demonstrates that the majority of that group (1,440 providers) never 'take' more than five books. This is consistent with the assumption that we are looking at individuals that search for specific titles, instead of those who are downloading as much as possible.

The ratio of nationalities is next to be examined. The percentages of all visitors of the online library can be used as a benchmark to compare against the clusters. A cluster containing a considerable difference in nationalities combined with a substantial difference in the range of subjects might signal that the books and providers have a meaningful connection. To enable this, we need to list the nationalities of all providers. However, the data contains over 160 different countries, ranging from Albania to Zimbabwe. The goal is to find significant differences, not a complete list. Therefore, the benchmark can be simplified to the ten countries with the highest usage. When the 'top-10' of a cluster contains countries not in listed in the benchmark, this is a clear – and easy to detect – signal.

6.3.4 The influence of the collection

The previous sections discussed the choices made to quantify the most useful aspects of the publications in the collection, and its users. These aspects are used to analyse how the readers – through the providers – interact with the books in the digital library. In other words: the collection shapes the possible actions of the readers. This leads to the question whether changes in the collection lead to changes in usage. To test this, the same investigation is carried out using data from 2014, two years after the first analysis. During that period, the collection of the OAPEN Library doubled from just over 1,100 titles to more than 2,300 titles. This growth influenced the collection on both axes: subject and language.

The growth of the collection altered the ranking of the subject categories and the languages. In 2012 "A - The arts" ranked third and "K - Economics, finance, business & management" ranked fourth. In 2014, this was reversed. The same holds true for "P - Mathematics & science" – ranked sixth in 2012 and seventh in 2014 – and "D - Literature & literary studies" – ranked seventh in 2012 and sixth in 2014. Within the languages, the ranking of Danish changed from fifth to seventh. Furthermore, due to the influx of titles in English, German and Dutch, the percentage of Italian language titles plummeted from 11% in 2012 to 5% in 2014.

The differences between 2012 and 2014 indicate that the focus of the collection may have shifted. Does this also lead to differences in usage?

6.4 Analysis

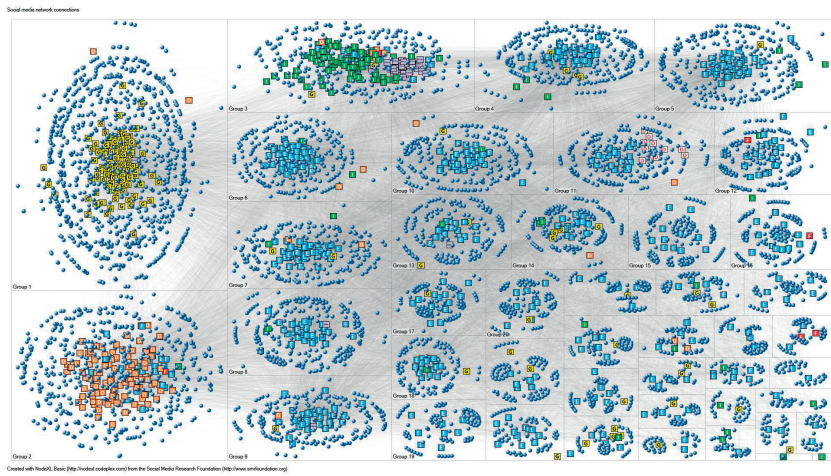
The previous section discussed the way the book and provider data was quantified: what aspects are to be examined? Furthermore, the collection of the OAPEN Library has expanded extensively, which also affects the usage. In this section, the analysis of the collection's usage in 2012 and 2014 are presented, exposing the difference in download patterns.

6.4.1 Examining clusters – the OAPEN collection in 2012

The first step towards answering the questions discussed in the previous sections is examining the usage patterns that occurred in 2012. In section 6.4.3, book downloads in 2014 are examined and the differences will be discussed. During the examined period – lasting three months – 967

different titles were downloaded by 5,180 providers. The total number of downloads is 34,345.

Figure 2 Clustered providers and books in the OAPEN Library, 2012



The linked titles and providers are clustered using the Wakita-Tsurumi (2007) algorithm, resulting in 43 clusters ranging in size from 1,000 elements (125 books and 875 providers) to a cluster consisting of exactly one book and one provider. And so we need to consider the number of clusters to investigate. At what point is the cluster too small to convey meaningful results? As this kind of study is scarce, there are no tried and tested guidelines. This paper's result might be considered to be a proof of concept, where the additional question of the optimal number of clusters is ignored for now. Instead, the – somewhat arbitrary – boundary is set at the ten largest clusters.

It must be noted that the data used in the Wakita-Tsurumi algorithm consists of nothing more than a unique code for each book, and the name of the provider. For instance, the connections between "uni-mannheim.de" and "422333"; "uni-mannheim.de" and "391039" are part of the data. After deploying the algorithm, the provider "uni-mannheim.de" is classified as German, and the books are identified as *Vernetztes Leben. Soziale und digitale Kulturen* and *The Practices of Happiness : Political Economy, Religion and Wellbeing*. Thus, the algorithm cannot be influenced by aspects of the providers or the books.

The data of each cluster has been analysed based on the following procedure. Firstly, the ranked subject classifications and the languages of the books in each cluster are compared to the complete collections' data. The

second question is whether the cluster's providers division in nationalities are in line with the percentages for the complete set. Substantial changes trigger a further examination of the book's subject by assessing keywords and titles.

The analysis resulted in the following 'named clusters':

- Cluster 1. German language books. Books in the German language, mostly downloaded by readers from Germany, Austria and Switzerland.
- Cluster 2. Dutch language books. Books in the Dutch language, mostly downloaded by readers from The Netherlands and Belgium.
- Cluster 3. Italian language books. A majority of the books is written in the Italian language, mostly downloaded by readers from Italy.
- Cluster 5. Film and Media. Books in the English language. The cluster contains a large group of books discussing film studies, plus a few titles on media or theatre studies.
- Cluster 6. Migration. Books in the English language, focused on migration studies.
- Cluster 9. Indonesia and South-East Asia. Books in the English language, mostly discussing Indonesia, in combination with works on South East Asia.

The appendix and the complete data set are available here: http://dx.doi.org/10.17026/dans-x72-d9h2.

6.4.2 Analysis results – 2012

Table 1 OAPEN Library: cluster analysis results (2012)

Cluster	Title	Books: Subject classifications	Books: Language	Readers: Nationality	Books: Keywords
1	German Language books	L – Law ranked #1, compared to #6. Most books on Law are written in German.	97% German, compared to 21% in the data set as a whole.	Germany ranked #1 (65%), #2 Austria (10%), #3 Switzerland (8%). Both Austria and Switzerland are not part of total collection top 10.	
2	Dutch Language books	D - Literature & literary studies ranked #3 compared to #7.	81% Dutch, compared to 11% in the data set as a whole.	Netherlands ranked #1 (64%), Belgium ranked #2 (12%). Belgium is not part of the total collection top 10.	

Cluster	Title	Books: Subject classifications	Books: Language	Readers: Nationality	Books: Keywords
3	Italian language books	G - Reference, information & interdisciplinary subjects ranked #3, compared to #8.	64% Italian, compared to 10% in the data set as a whole.	Italy ranked #1 (44%).	
5	Film and Media	A – The arts ranked #1, compared to #3.	92% English, compared to 52% in the data set as a whole.	USA ranked #1 (42%), #2 Great Britain (13%).	27 of the 40 titles discuss film studies.
6	Migration	J - Society & social sciences ranked #1 (64 % compared to 31 %).	88% English, compared to 52% in the data set as a whole.	USA ranked #1 (22%), #2 France (13%) compared to #6, #3 Spain (12%). Spain is not part of the total collection top 10.	38 of the 47 titles discuss migration.
9	Indonesia and South-East Asia	J - Society & social sciences ranked #1 (45 % compared to 31 %).	92% English, compared to 52% in the data set as a whole.	USA ranked #1 (32%), #2 Indonesia (18%), #3 Australia (9%) compared to #10. Indonesia is not part of the total collection top 10.	22 of the 35 titles discuss Indonesia or South-East Asia.

The role of nationality and language is visible in the largest clusters. The first cluster contains 125 books, and 122 of those titles are in the German language. Also, the 'top 3' nationalities of the providers are German speaking countries. The same holds true for the second cluster, which consists of a large majority of Dutch language monographs. Here, the Dutch and Belgian providers rank one and two, respectively. And the third cluster is dominated by Italian languages monographs and Italian providers. Within these clusters, the ranking of the subjects seems to reflect the division of the books in the respective language. For instance, in Cluster 1, Law ranks first. The explanation can be found in the relative large number of law titles in the German language.

Within the cluster on Film and Media, English plays a major role. As is the case with Cluster 6 and Cluster 9, the USA providers are now ranked at the first place. In these three clusters, subject plays a major role. This can be inferred from the differences between the classification within the clusters compared to the whole collection, and an examination of the keywords and

titles of the books. We might conclude that the role of English is different from German, Dutch or Italian: it is not a defining property of a cluster.

In the case of Cluster 9 – Indonesia and South-East Asia – the interest through Indonesian and Austrian providers can easily be explained by a regional focus. In contrast, the international usage of the clusters on Film and Media or Migration do not show such a clear pattern. In the case of Cluster 6 – Migration – the spread of providers is relatively even: there are no countries with a much stronger interest compared to the other 'members' of the cluster. It is noteworthy however, that the "providers top 10" lists Spain, Greece, Austria and Hungary. All of these countries are not part of the total collection's top 10. Yet, in these countries – and also in France, Poland and Germany – immigration is a widely-debated topic. This might point to a regional interest, but the signal is not as strong compared to the data by Cluster 9. See the Appendix for more details.

6.4.3 Examining clusters – the OAPEN collection in 2014

When the same method is applied to the data of a three-month period in 2014, the differences are striking. During that time, 2,334 different titles were downloaded 60,238 times, roughly twice the amount of 2012. However, the number of providers 'only' raised 20% to 6,316 providers. Most of these providers (69%) downloaded one book in a month; and the total percentage of providers that downloaded 5 titles or less is 98%. Furthermore, the 'country top ten' list contains the same countries, with the exception of Ukraine, which replaces Poland.

The question is whether the changes in the collection affected the usage: is it possible to detect the same clusters? Here, the number of clusters is comparable to 2012: 41. The largest cluster contains 244 books and 723 providers, while the smallest cluster consists of one book and two providers. Again, the largest ten clusters are compared to the data of the complete collection.

The analysis resulted in the following 'named clusters':

- Cluster 1. German Language books. Books in the German language, mostly downloaded by readers from Germany. Providers from Austria and Switzerland are ranked third and fourth. Ranked #2 are providers from the US.
- Cluster 2. Dutch Language books. Books in the Dutch language, mostly downloaded by readers from The Netherlands. Comparable to Cluster 1, the US providers rank second, followed by Belgium providers.

- Cluster 3. The largest clustering of Italian language books, but this cluster also contains a large portion of German books. Here, Italian rank first, followed by German providers.
- Cluster 5. Indonesia and South-East Asia. This cluster is comparable to Cluster 9 of the 2012 data, containing books in the English language, mostly discussing Indonesia and South-East Asia.
- Cluster 9. Australia and the Pacific region. English language books on subjects related to Australia and the Pacific region. The US providers are ranked first, Australian second.

6.4.4 Analysis results – 2014

Table 2 OAPEN Library: analysis results (2014)

Cluster	Title	Books: Subject classifications	Books: Language	Readers: Nationality	Books: Keywords
1	German Language books	R - Earth sciences, geography, environment, planning #4, compared to #10. Most books on this subject are written in German.	91% German, compared to 24% in the data set as a whole.	Germany ranked #1 (46%), #3 Austria (10%), #4 Switzerland (6%). Both Austria and Switzerland are not part of total collection top 10.	
2	Dutch Language books	Consistent to the number of titles in Dutch, K - Economics, finance, business & management#2 and H - Humanities #3	92% Dutch, compared to 19% in the data set as a whole.	Netherlands ranked #1 (52%), Belgium ranked #3 (9%). Belgium is not part of the total collection top 10.	
3	Italian language books		23% Italian, compared to 5% in the data set as a whole. However, 43% of the books are German.	Italy ranked #1 (22%), #2 Germany (20%).	

Cluster	Title	Books: Subject classifications	Books: Language	Readers: Nationality	Books: Keywords
5	Indonesia and South-East Asia	J - Society & social sciences ranked #1 (43 % compared to 31 %). K - Economics, finance, business & management ranked #2 (16% compared to 8%)	98% English, compared to 47% in the data set as a whole.	Indonesia ranked #4, #5 India, #6 Pakistan. Indonesia, India nor Pakistan are part of the total collection top 10	60 of the 119 titles discuss Indonesia or South-East Asia.
9	Australia and the Pacific region		70% English, compared to 47% in the data set as a whole.	Australia ranked #2 (22%) compared to #10.	74 of the 124 titles discuss Australia or the Pacific region.

The three largest clusters are once again connected to books in a specific language, without a specific emphasis on a subject. We could argue that the contents of Cluster 3 are relatively 'diluted': the number of German books is higher than the books in Italian. However, it contains the largest concentration of Italian monographs, combined with a large Italian readership.

It is noteworthy that both Film studies and Immigration are less visible, while books focusing on the Oceania region are easily spotted. An explanation may be found in the influx of new titles in the OAPEN Library. In 2013, the collection grew with over 300 titles published by ANU Press, part of Australian National University.

The number of titles on immigration did not grow as spectacular. Snijder (2013) discusses the dissemination of books by the IMISCOE Research Network on international migration, integration and social cohesion. The majority of those book made available through the OAPEN library in 2012, and the data of 2012 contains 50 IMISCOE titles. Most of them – 34 books – are found in cluster 6: Immigration. Between 2012 and 2014, only ten more titles were added – a total of 60 books. Compared to the growth of the complete OAPEN Library collection, this is a modest increase.

The role of American providers is also unmistakable. According to *The World Factbook* (Central Intelligence Agency, n.d.), the number of American internet hosts is 505,000,000. A large number, compared to the second country on the list – Japan – which contains 64,453,000 hosts; a factor 7 less. Given these amounts, the prominent role of US providers is not surprising.

6.5 Creating recommendations based on clusters

We have discussed before that personal recommendation systems cannot be used in open access libraries. It is nonetheless possible to detect patterns in the use of the library, and with relative simple means, meaningful clusters of books and providers can be detected; the current results can be seen as a proof of concept. Contrary to the assertion of Lynch (2002), it is possible to identify – up to a point – which user communities will engage with the digital library. The detected patterns help at the very least to define interests by larger groups of readers; a precondition for the creation of new services.

A possible service could entail listing groups of titles, to be presented to certain groups of providers. The clustering results can be converted into a set of recommendation rules, based on the contents of the book combined with the nationality of the provider. For instance, the results from cluster 1 of section *6.4.2 Analysis results – 2012* could be transformed into the following 'recommendation rule': If the provider is based in Germany, Austria or Switzerland and has downloaded a book in German, present a list of all German books. Likewise, this combination of provider nationality and non-English books could be applied to cluster 2 – Dutch language books – and cluster 3 – Italian language books.

There are also subject-based clusters, for instance cluster 5 in section *6.4.4 Analysis results – 2014*. Here, the recommendation might run along the lines of presenting English language books on Indonesia or South-East Asia to providers based in Indonesia, India or Pakistan. Cluster 5 of section *6.4.2 Analysis results – 2012* would lead to a more generic rule: if one English language book on film and media studies has been downloaded, present all English language titles on this subject.

The suggestions listed are not the result of a careful curation by a librarian, but are purely based on the usage patterns that have been uncovered. Recommender systems are based on the preferences of individuals; the suggestions here are based on the preferences of "implicitly defined communities" as described by Kumar *et al.* (1999). In this way, the flexibility of recommender systems is deployed, without violating the privacy of individuals.

6.6 Discussion

In the previous sections, we have seen the analysis and the possible recommendations based on its results. Yet, on a more abstract level there are

several other aspects to reflect upon: the role of regional interests and how well the deployed algorithm performs on the total collection.

When the different clusters are analysed, the influence of language communities is profound. It might not come as a surprise that readers in languages other than English tend to be more interested in publications written in their 'local' language. Thus, the clusters of books in German, Dutch or Italian are read mostly by native speakers. The language effect is quite strong: within these clusters it is hard to find a subject based focus. In contrast, if publications in English are taken into account it is still possible to find clusters whose subject is closely tied to a region. This is especially visible in the clusters focused on Indonesia and Sea East Asia, and the cluster concerning Australia and the Pacific region. Even the subject of immigration could be seen as a regional – mostly European – concern.

One might argue that the available data tends to point in this direction: the main thing known about the readers are their provider's countries. Furthermore, one of the aspects analysed is the distribution of nationalities. Given this procedure, it is rather hard to miss 'regional' patterns. On the other hand, region is not the only scrutinised aspect. The books' subject and language are also taken into consideration. As an additional test, all twenty clusters have been analysed using subject and language only. This did not lead to new insights.

The clusters are the results of algorithms – predefined procedures. Deploying these procedures leads to interesting results: uncovering usage patterns. We also saw that the results differ: in 2014, no clusters concerning Film and media or Immigration were detected. Instead, the large influx of books by an Australian publisher was visible. Does this mean that interest in film or immigration studies has diminished? That might be possible, but another option seems more likely: the performance of the algorithm when it is applied to the collection.

Between 2012 and 2014, the collection nearly doubled. As was illustrated by the example of the IMISCOE series and the books in Italian, the number of books concerning a certain subject may not always keep pace with the collection's growth. The algorithm only detects the 'strongest' patterns, based on relatively large groups. Thus, smaller clusters of books and readers may go unnoticed.

The background section discussed several types of recommender systems. The variety hints at room for improvement: there is no single best solution. This may also apply to this paper's procedure; other procedures and algorithms may yield improved results. A recent paper by Gläser, Glänzel, & Scharnhorst (2017) illustrates this: the authors describe the search for

optimized deployment of algorithms to cluster articles into "thematic clusters". Different algorithms lead to different results, all of which might be valid in their own way. The theme of this paper is also a clustering problem, and thus the results by Gläser *et al.* could be applied here as well.

Simply put: the question is how to proceed from this starting point in order to create a fully functioning system? There are several points to explore. Firstly, the results of several clustering algorithms should be evaluated. We have seen that the currently used algorithm detected other groups in the collection data of 2012 and 2014. Will other algorithms lead to strongly differing results? Another avenue to explore is recursive use: deploying the algorithm again on the clusters, in order to find 'sub groups'. Earlier in the paper, the question which clusters should be investigated was mentioned. This might be an additional study. Lastly, the current analysis is depending on human judgment, especially on the book's subjects. In an open access library, the documents are available in a full text form. Using text mining techniques might help to automatically cluster the books, based on common words or word sequences. It would be interesting to see if these 'subject clusters' overlap with the clusters of providers.

6.7 Conclusion

This paper attempts to unravel the paradox of open access libraries: created for maximum dissemination, but deploying one of the most powerful tools to support its users leads to questions about privacy. Recommender systems are used widely and with great success, but are built on storing information about individuals. This is hard to accept from a privacy point of view, and open access libraries are not normally equipped to individually track their readers. However, every library functions better when it understands the needs of its patrons.

Open access libraries are web based by definition, and the usage through providers indicates the level of interest for each document. The thousands of data points require the use of automated procedures. Applying social networking analysis techniques helps to uncover patterns of usage that are very hard to spot in a different way. With relative ease, it is possible to run a meaningful analysis of the interests of groups of readers.

This paper's results can be seen as a proof of concept; a possible starting point for recommendations built on usage that retain the privacy of individual readers.

6.8 Acknowledgements

The author would like to thank Professor Paul Wouters of the Centre for Science and Technology Studies (CWTS) and Professor Frank Huysmans of University of Amsterdam for commenting on the draft version of this paper.

7 Measuring monographs : A quantitative method to assess scientific impact and societal relevance

Snijder, R. (2013). Measuring monographs: A quantitative method to assess scientific impact and societal relevance. *First Monday*, 18(5). https://doi.org/10.5210/fm.v18i5.4250

7.1 Monographs under pressure

In the Humanities and Social Sciences (HSS), monographs – instead of articles – play an important part in communicating scholarly results.[1] However, the publication of (paper) monographs faces challenges. Greco and Wharton describe the problems faced by university presses, resulting in smaller print runs per title and declining sales to libraries and institutions (Greco & Wharton, 2008). Also, Thomson describes falling print runs and declining sales (Thompson, 2005). The decline in dissemination of scientific monographs is further illustrated by the Association of Research Libraries. The expenditure for journals grew from more than $1,400,000 in 1986 to over $7,513,000 in 2011. This contrasts sharply to the $1,120,000 spent in 1986 and $1,936,000 spent in 2011 on monographs("Association of Research Libraries (ARL) :: ARL Statistics 2009-10," 2012). Williams *et al.* also describe a decline of sales combined with negative effects on print runs, but conclude that the monograph remains the single most valued means of scholarly publishing within the field of Arts & Humanities (Williams *et al.*, 2009). Withey *et al.* conclude that the economic model supporting monographs depends for a significant amount on subsidies (Withey *et al.*, 2011). This funding model can only be sustained if the return on investments is clear.

This raises the question why monographs are used more than journal articles. The answer might be found in the definition by Chodorow: "The monograph is a large, specialized work of scholarship that treats a narrow

1 Psychology is an exception: in this field articles are used more than monographs (Schaffer, 2004).

topic in great detail." He adds that "monographs are principally about establishing facts or narrative in a set of fields in which facts and narratives are often hard to establish" (Chodorow, 1999). Due to its size, a monograph enables researchers to describe the results of research spanning a long period in sufficient detail. It is therefore best suited for the type of research mostly conducted in the field of HSS. It is targeted at a specialised audience, in contrast to a 'text book' which is designed for a more general audience. However, in this article we will see that there is an interest in monographs by the 'general public'.

The monograph clearly performs a useful function in the field of HSS, especially because of its length. An example of scholarly use of monographs is described by Mendez and Chapman who investigated the role of monographs as sources in the field of Latin American History. They conclude that the use of monographs as secondary sources – after a decline in the period 1985 to 1995 – is elevated to a higher level in 2005 (Mendez & Chapman, 2006).

However, scholars in the Humanities and Social Sciences are expected to describe their contribution to society. As in the field of Science, Technology and Medicine (STM), there is a need to assess the value of scholarly output.

7.2 Scientific impact, societal relevance and monographs

In several countries government policies have been developed to assess the quality of scientific and scholarly research, in other countries the assessment is done by academies of sciences. The aim is to enhance the quality of scientific work and to maximise the societal benefits deriving from it. Assessing the quality of research is normally done on two levels: at the level of individual scientists or scholars and at the level of scientific or scholarly output. The first level is measured through 'esteem indicators' as prizes and scholarly positions, or the amount of international influence. At the level of output we find 'internal assessments': peer review of documents and 'external assessments' through bibliometric indicators, such as high ranking journals, book series or publishers (Royal Netherlands Academy of Arts and Sciences, 2010). Furthermore, the assessment must take into account the variety of output forms – it should not be limited to journal articles – and the bureaucratic burden must be limited.

On top of this, research and its outcomes can be categorised as Mode 1 and Mode 2, where Mode 1 research is done within the academic discipline, and Mode 2 research aims at the application of research outcomes. This

concept was introduced by Gibbons (Gibbons, 1994); the application in research evaluation is recently discussed by Ernø-Kjølhede and Hansson (Ernø-Kjølhede & Hansson, 2011). Leydesdorff and Etzkowitz use a different angle by looking at the relations between universities, governments and industries: the "Triple Helix" (Leydesdorff & Etzkowitz, 1996).

Creating the best possible scientific or scholarly output is not a goal in itself; the output should be used by others. Usage by scientists is termed scientific impact; usage by others is termed societal relevance. Usage is not exactly the same as impact; it functions as an *indicator* for impact. Measuring scientific impact in the field of HSS is poorly developed compared to the field of STM. In the field of STM, the use of bibliometric measures such the Journal Impact Factor (JIF) or the h-index is often discussed, although its application is controversial and often inappropriate. In the field of HSS – where articles play a smaller role in disseminating research results – similar tools are not widely available.

However, Nederhof and Linmans have discussed the usage of bibliometric tools in the Humanities and the Social Sciences. Nederhof investigated the possibilities of bibliometric research in the field of HSS and concludes that it is possible to use the same methods as deployed in STM. It could be done if more types of publications – monographs and journals not covered by ISI – are taken into account and by applying impact indicators that compensate for the smaller volumes of citations in the humanities and social sciences, compared to the field of STM (Nederhof, 2006). Linmans focuses on citations per author, not from a certain period but on lifelong citation data. This method aims to make more citation data available, which should lead to more robust results (Linmans, 2009).

Alternatives to the 'standard' bibliometric methods have also been described. White *et al.* discuss 'libcitations', where the number of academic libraries holding a certain book is the unit of measure. The collection of a library is formed based on qualitative decisions; a monograph that is acquired by a large number of libraries is 'better' than a monograph that only resides in a few libraries (White *et al.*, 2009). The MESUR project is not only based on counting citations, but also focuses on the usage of online sources – mostly journal articles – by scientists. The authors see online usage as a better indicator for scientific impact than citations (Bollen *et al.*, 2009; Bollen, Van de Sompel, & Rodriguez, 2008). The method described in this article is also based on measuring online usage, but here the focus is not on journal articles; it is on monographs instead. Online usage is also discussed by Herb *et al.*, publishing work on the usage and interface design of repositories – the most widely used way to disseminate open access

documents (Herb, 2010; Herb, Kranz, Leidinger, & Mittelsdorf, 2010). While the discussed research uses quite different modes of operation, all of it is aimed at scientific impact, not on societal relevance.

In order to measure the usage of scientific or scholarly output in society, more elaborate methods are needed. Several researchers have published work on defining societal relevance and the evaluation of the current frameworks. The methodology described by Lyall encompasses focus groups, questionnaires, desk research and stakeholder analysis; a method which does not seem to minimize bureaucratic demands (Lyall, Bruce, Firn, Firn, & Tait, 2004). In the Netherlands, the same methodology was presented by the QANU organisation (Bennink, Meijer, Wamelink, & Zuijdam, 2008). The SIAMPI project defined three types of indicators (termed 'productive interactions'): direct or personal interactions; indirect interactions through texts or artefacts and financial interactions through money or 'in kind' contributions (Spaapen & van Drooge, 2011). The method described here measures one of the interactions: through the texts of electronic version of monographs. Furthermore, current policy programs aimed on societal relevance are studied. An example is the case study by Grant *et al.* of the Australian RQF, the UK RAISS method, the US PART framework and the Dutch ERiC framework (Grant, Brutscher, Guthrie, Butler, & Wooding, 2010).

Very little is known about the societal relevance of monographs. Only recent, Serenko *et al.* have published research on societal relevance in the field of knowledge management and intellectual capital (Serenko, Bontis, & Moshonsky, 2011). Within knowledge management, there is a relatively clear distinction between scholars and practitioners. As all stakeholders are known, the flow of knowledge from one group to the other is not hard to follow. In the social sciences, government agencies are considered to be a major benefactor of the scientific results. Several usage studies – primarily based on surveys and interviews – have been published (Bell, Shaw, & Boaz, 2011; Landry, Amara, & Lamari, 2001; Landry, Lamari, & Amara, 2003). In other disciplines in the humanities, the picture is less clear. Benneworth and Jongbloed show that the stakeholders – in other words: the groups that would primarily benefit from research – are less visible to universities (Benneworth & Jongbloed, 2009). Of course, if stakeholders are not known, it is impossible to perform the kind of qualitative research described by Lyall.

Measuring scholarly impact and societal relevance in the humanities and social sciences is not without problems. When methods based on bibliographic data are used to assess scholarly impact, the lack of data makes the results less reliable. The proposed and used methods to assess societal influence are labour intensive; this requires a large investment in

time and money. Furthermore, the results are dependent on self-assessment of the respondents. Of course, this may introduce bias: depending on the respondent the perceived results may be too positive or too negative. In the case of humanities, the picture becomes even less clear due to uncertainties about the stakeholders.

This article describes a method that may complement the current research on scientific impact and societal relevance. It relies on analysing data generated by usage of electronic versions of monographs. Every time a reader opens a web page or downloads a document, information about the organisation through which the reader accesses the web is recorded. By assessing this information, it is possible to determine the type of organisation and the county of origin. Due to extensive use of automated tools it is less labour intensive than the previously described methods, and it may uncover groups of users, even in disciplines where stakeholders are not well known. The method is tested on data generated from the OAPEN Library.

7.3 The method

The method is based on the fact that books can be made available online, in full or partial, through a dissemination channel. Those channels may impose restrictions such as full or limited availability, enabling downloading, printing etc. Examples of dissemination channels are the Google Book Search program, institutional repositories or e-book collections of academic libraries. Each of these channels collect usage data, such as the number of views or downloads and some information about the user. Almost all web based channels list the web address of the 'provider': the organisation that grants access to the internet. So, if a researcher of Leiden University downloads a book using her or his office equipment, the web address (www.leidenuniv.nl) of that university will be logged. Basic information such as address and telephone number are publicly available and can be found using the so called 'WHOIS protocol' ("WHOIS - Wikipedia," n.d.). By combining the usage data and information about the provider, we can make an assumption about who is using a specific monograph. To put if differently: the type of provider is used to assess the type of reader. In the example used, the reader is affiliated with an academic institution, based in the Netherlands.

7.3.1 Defining stakeholders: scientific impact and societal relevance

If the dissemination channel is open to everybody, it may attract users from all kinds of organisations. Not everybody will have an academic organisation as provider; it may be another type of organisation or it will be an Internet Service Provider (ISP). It then becomes necessary to define several groups of organisations. Here, the following categories are used: academic; government; business; non-profit organisations and the general public. Academic users are seen as the main audience for monographs. Based on the literature on societal relevance, we could divide the other types of readers of monographs into the following categories: government, business and general public. If the provider is an ISP, the reader cannot be linked to an organisation. This could mean that the reader is not acting as a member of an organisation, and may be categorised as a member of the general public. In this article, another type of organisation is proposed: non-profit organisations.

Within the humanities and social sciences, we might expect to find stakeholders that are not commercial, who play a role in the discipline. In the social sciences, government is seen as a significant stakeholder, and government policies regarding certain subjects – for instance: immigration, environment – receive considerable attention from non-profit organisations. Societal relevance by those types of organisations is therefore also to be expected. As discussed before, the situation in the humanities is less clear and stakeholders are not identified. Still, we might expect usage from non-profit organisations. For instance, national history may cause considerable interest.

Apart from the provider, information about the country from which the data request originated is available, indicating the nationality of the reader. This information can be used to classify the usage a bit further: national versus international. In order to classify usage to be national or international, we need to establish the 'nationality' of a monograph. Several choices are available: the nationality of the author(s), the country of the author's organisation or the country of publication. Here, the country of publication is used; the information about authors or their organisations was not available.

This method can be used to measure the scientific impact and the societal relevance of one monograph. The ratio of academic readers versus other users may be used as an indication of the level of scientific impact and societal relevance. Examining a group of monographs enables us to look at other aspects as well: what is the influence of the monograph's

subject or is language a barrier for international usage? The amount of national and international usage could be closely linked to the language of the monograph. When looking at the monograph's subject, it may be possible that different scientific disciplines display other usage patterns. For instance, the percentage of users connected to a government organisation may be larger in the social sciences than in the humanities.

Most literature on societal relevance does not explicitly focus on international usage; from a policy point of view societal relevance is looked at on a national level. Policy makers are more likely to prioritize usage on a national level as a way to measure the return on investments in science done by national governments. Still, the international usage should also be taken into account. As discussed before, international usage is used as an indication of esteem. The percentage of usage outside national borders may give an indication of the importance of the work. This reflects on the authors; one of the 'esteem indicators' is the level of international interest.

The used data set contains books that are published in West European countries. Usage is global however, ranging from Albania to Zimbabwe. This also includes the so-called "developing countries", with more limited financial resources. The digital divide between the developing countries and the developed countries could be described as a financial barrier to access (Swan & Hall, 2010). Here, all monographs used are published in open access, therefore this barrier does not exist here and this aspect will not be discussed in this article.

Conclusions regarding these statistics must be drawn with caution. First of all, the information found using the WHOIS protocol must be interpreted: what type of organisation is described? If the organisation is a university, it is quite clear. The question where to draw the line between an ISP and another type of commercial organisation is less easy to answer. Also, organisational affiliation does not tell anything about professional roles. For instance, if the provider is a university, there is no way to tell whether the reader is a student or a professor. Likewise, if the provider is an ISP, we cannot be sure the reader used the online monograph for personal or professional reasons. Regarding nationality, this too is not a 100% match: one could easily imagine a Spanish reader downloading a monograph while in the USA. The user statistic would then indicate the USA as country of origin. A possible remedy could be found in using a survey, asking readers about their professional affiliation, role and nationality. And finally, here we measure the number of downloads. The number of downloads is an indication of readership: we can assume that the more a book has been downloaded, the more is has

been read. But we cannot state that 100 downloads equal 100 people reading the book cover to cover.

7.3.2 Selecting a channel to measure usage

In order to measure the usage of electronic monographs, we need access to dissemination channels. One may consider academic libraries to be the obvious choice. However, there are certain drawbacks to this dissemination channel. First of all, measuring usage from an academic library constricts the user population to the staff and students of that particular academic institution. The composition of the group needs to be taken into account. For instance, if faculties are significant different in size, it may reflect on the usage measured. A far more serious problem is the fact that academic libraries are not open to outsiders, making it impossible to measure societal relevance. Furthermore, usage 'outside' of the library catalogue – of monographs found through search engines – is not measured.

Collections of monographs are not only found in libraries. Academic publishers also have access to dissemination channels. Publishers have a different interest from academic libraries; instead of serving one academic community, publishers need to be known as widely as possible. This reflects on their usage of dissemination channels: at the very least, information on all available publications are accessible to everybody. Therefore, usage data is not restricted to certain groups and could be used to measure both scientific impact and societal relevance. Furthermore, access to the data is not channelled through a library catalogue, but is wide open to both search engines and other linking mechanisms – such as the Facebook website (Vascellaro, 2009).

7.4 The OAPEN Library as dissemination channel

The method was tested on the OAPEN Library, which was officially launched in September 2010. The OAPEN Consortium describes it as "an Online Library containing a freely available, quality-proven and multilingual collection of monographs from various fields of HSS" (OAPEN Consortium, 2011). It is a web based collection of monographs, which are all available in open access. The website offers several ways to make its contents accessible: it enables searching and browsing, readers can share book descriptions via social media and it contains several data feeds (Open Access Publishing in European Networks, 2010b).

The OAPEN Library was used because its collection contains a diverse range of subjects, published by dozens of publishers and in several languages. This creates a large data set, which contains sufficient large sets of monographs with the same language, subject etc. For this article, the number of downloads of the full year 2011 as measured through the Google Analytics program were used. Google Analytics only measures the number of downloads that result from a visit to the OAPEN Library website. This does not draw a complete picture; all monographs can be directly downloaded, without browsing the OAPEN Library website. So, if a reader uses a search engine such as Google or Bing to find a book and downloads it directly from there, the download will not be registered in Google Analytics. The total number of downloads in 2011 is larger than 300,000. At this moment, not all user statistics are available. Therefore, the Google Analytics data will be used.

The data set used consists of a diverse set of monographs: 859 titles, published by 30 publishers. There is also a wide range of languages available: Danish; Dutch; English; French; German; Italian; Latin; Norwegian; Spanish and Welsh. For those titles, 25405 downloads were measured, by 1574 unique providers. Each provider was classified as one of the following types: academic, government, business, non-profit or ISP. For each download, the provider was further classified as national or international, depending on the county of publication and the country of the provider: if the country of publication equals the country of the provider, the provider is national; otherwise it is classified as international. So, if the University of Exeter downloads a book by the Dutch publisher Brill, it is classified as Academic (International). When a book published by Manchester University Press is downloaded by the University of Exeter, it is classified as Academic (National).

7.5 Setup of the research

The goal of the research is to test the method and gather quantitative data about scientific impact and societal relevance of scientific monographs. This type of research is new; therefore, no best practice is established. Here, the percentage of downloads per type of provider is used as a measure for scientific impact and societal relevance, combined with the average number of downloads per group of titles. By comparing these groups, we may be able to find significant differences. No benchmark is available, so it is not possible to say how well a certain monograph 'performs'.

Monograph usage can be measured on two levels:
1. At the level of separate titles
2. At the level of the complete collection

7.5.1 Measuring usage at the level of separate titles

In this article, the data at the level of the complete collection or at the level of large subsets will be discussed in most detail. It is possible to analyse each monograph's usage. The following example shows the usage data for the book *Globalization contested: An international political economy of work*[2] written by Louise Amoore and published by Manchester University Press in 2002.

About 23% of the usage comes from (international) academic institutions, and almost 70% is generated by foreign ISPs. The remaining usage is generated by a company, a British ISP and a non-profit organisation: the International Atomic Energy Agency. We will see that those figures are no exception: the average usage percentages per provider type are more or less along these lines. The international usage is truly worldwide; this is also typical for all the measured data, which originated from 102 countries.

Table 1 Usage data of one book

Organisation	Type	Country	Downloads
University of Queensland	Academic (International)	Australia	1
University of Hong Kong	Academic (International)	China	1
Universität Duisburg-Essen	Academic (International)	Germany	1
University of the Aegean	Academic (International)	Greece	1
Hokkaido University	Academic (International)	Japan	1
Universiteit van Amsterdam	Academic (International)	Netherlands	1
Universidade do Porto	Academic (International)	Portugal	1
National University of Singapore	Academic (International)	Singapore	3
Webtrade Ltd.	Business (International)	Ireland	1

2 See: http://oapen.org/search?identifier=341340

Organisation	Type	Country	Downloads
International Atomic Energy Agency	Non-profit (International)	Austria	1
Virgin Media	ISP (National)	Great Britain	1
Belgacom	ISP (International)	Belgium	1
Telecel S.A.	ISP (International)	Bolivia	1
Cambodian ISP, Country Wide, Wireless IAP	ISP (International)	Cambodia	1
Ezecom	ISP (International)	Cambodia	1
Bell	ISP (International)	Canada	1
Cytanet	ISP (International)	Cyprus	1
UPC Broadband	ISP (International)	Czech Republic	1
Arcor AG	ISP (International)	Germany	1
Ewe Tel	ISP (International)	Germany	1
OTEnet S.A.	ISP (International)	Greece	2
Videsh Sanchar Nigam	ISP (International)	India	1
PT Telekomunikasi Indonesia, Tbk	ISP (International)	Indonesia	5
PT. Global Media Teknologi	ISP (International)	Indonesia	1
XS4All	ISP (International)	Netherlands	2
Ar Telecom	ISP (International)	Portugal	1
Astral	ISP (International)	Romania	1
BEOTEL-AS BeotelNet-ISP	ISP (International)	Serbia and Montenegro	1
Telia	ISP (International)	Sweden	1
Asia Infonet	ISP (International)	Thailand	1
TOT Content Farm Network	ISP (International)	Thailand	1
Farlep-Odessa ISP	ISP (International)	Ukraine	1
GoDaddy.com	ISP (International)	USA	1
RoadRunner	ISP (International)	USA	1
SYNCHRONOSS TECHNOLOGIES	ISP (International)	USA	1

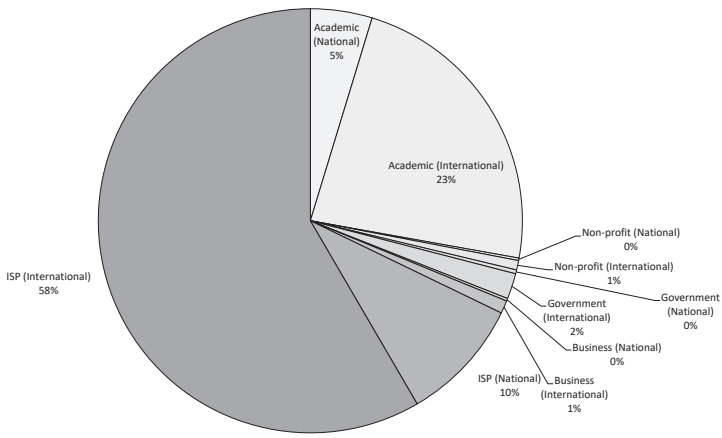
7.5.2 Measuring usage at the level of the complete collection

Looking at the usage data of all books, it is clear that most traffic comes from ISPs, followed by usage from academic institutions. While usage by government or business is discussed as the primary source of societal relevance, here it plays a minor role. Furthermore, 85% of the usage is international.

Table 2 Usage data of all books

Usage	Total	National	International
Academic	27.82%	4.71%	23.11%
Non-profit	0.91%	0.19%	0.72%
Government	2.25%	0.30%	1.95%
Business	1.18%	0.20%	0.98%
ISP	67.84%	9.44%	58.40%
Total		*14.84%*	*85.16%*

Figure 1 Downloads OAPEN Library



7.6 Are all ISPs equal?

The high percentage of usage coming from ISPs presents an unexpected problem. Without further refinement, almost 68% of the usage is hard to categorize. A method is needed to distinguish whether the usage comes from users whose organisation does not provide internet access or from users who are downloading the monographs 'from home'. The solution can be found by looking at the internet infrastructure per country, combined with the percentage of ISPs.

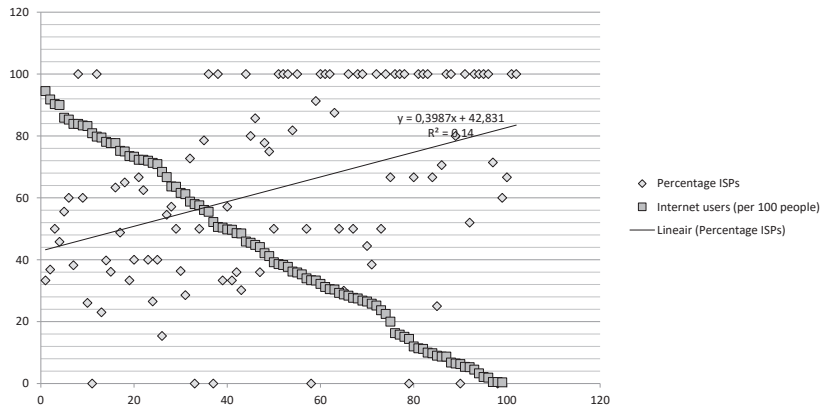
7.6.1 Internet infrastructure and ISPs

The internet infrastructure differs from country to country. We might assume that in countries with a highly-developed internet infrastructure,

most organisations are capable of directly providing internet access to their employees. In contrast, access to the internet will almost certainly be provided through an ISP in countries with a weakly developed internet infrastructure. In other words: we might expect that in countries with a highly-developed infrastructure, 'professional users' are more likely to use the internet access provided by their organisation and the users who access the OAPEN Library through an ISP are not doing that as part of their professional role.

In order to assess the state of the internet infrastructure per country, statistical data from the World Bank is used. The publication *The Little Data Book on Information and Communication Technology 2011* contains several indicators on the state of the IT infrastructure per country (World Bank, 2011). One of the indicators is the amount of internet users per 100 people. When this indicator is plotted against the percentage of ISPs per country found in the data, we find that in countries with a higher percentage of internet users – countries with a better developed infrastructure – the percentage ISPs is lower.

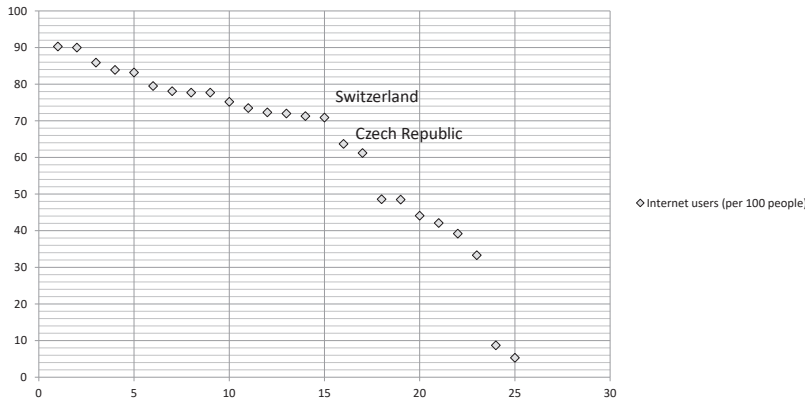
Figure 2 Percentage ISPs and internet users per country



When we look at the data we might assume that the people using a highly-developed internet infrastructure are less likely to use an ISP if they download books from the OAPEN Library in their professional role. So, downloads through an ISP from countries with a highly-developed internet infrastructure are more likely to be coming from non-professional users. The next question to answer is which countries are considered to have a highly-developed infrastructure. Plotting *all* countries displays a smooth descend from 94.5 internet users per 100 people (Iceland) to 0.5 (Ethiopia).

In order to find a suitable cut off point, the number of providers of all countries was listed. From this list, the 25 countries with the highest number of providers – regardless of the type – were selected, and the number of internet users per 100 people was plotted in the following chart.

Figure 3 Internet users (per 100 people)



The first cut off point can be found between Switzerland (70.9 internet users per 100 people) and the Czech Republic (63.7 internet users per 100 people). Therefore, it is assumed that all countries with 70 or more internet users per 100 people have a highly-developed internet infrastructure and ISP usage from these countries is more likely to come from the 'general public'.

The 25 countries with the highest number of providers are listed below.

Table 3 The 25 countries with the highest number of providers

Provider country	Total number of providers	Number of ISPs	Percentage ISPs	Internet users (per 100 people)
Sweden	36	18	50.00%	90.3
Netherlands	107	49	45.79%	90
Denmark	45	25	55.56%	85.9
Finland	34	13	38.24%	83.9
Great Britain	119	31	26.05%	83.2
Germany	113	26	23.01%	79.5
USA	181	72	39.78%	78.1
Canada	36	13	36.11%	77.7
Japan	30	19	63.33%	77.7
Belgium	41	20	48.78%	75.2
Austria	36	12	33.33%	73.5

Provider country	Total number of providers	Number of ISPs	Percentage ISPs	Internet users (per 100 people)
Poland	30	20	66.67%	72.3
Australia	45	18	40.00%	72
France	49	13	26.53%	71.3
Switzerland	25	10	40.00%	70.9
Czech Republic	42	24	57.14%	63.7
Spain	42	12	28.57%	61.2
Portugal	25	9	36.00%	48.6
Italy	53	16	30.19%	48.5
Greece	25	9	36.00%	44.1
Russia	63	49	77.78%	42.1
Brazil	24	12	50.00%	39.2
Ukraine	23	21	91.30%	33.3
Indonesia	51	36	70.59%	8.7
India	25	13	52.00%	5.3

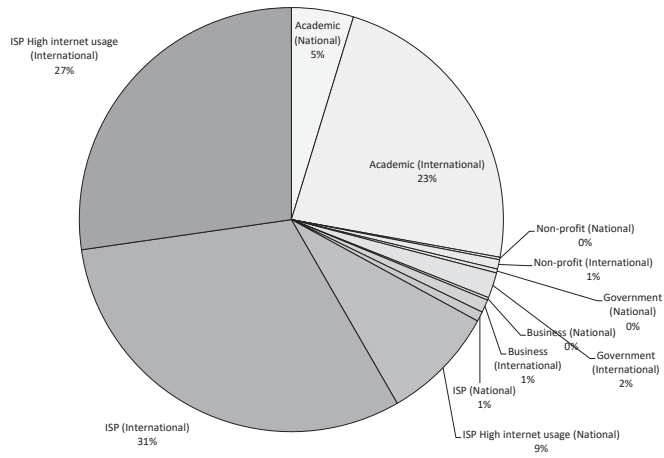
7.6.2 A refined categorisation of ISP usage statistics

Refining the categorisation of the ISP usage statistics does paint quite a different picture. The percentage of data generated by ISPs is now divided into 31.86% that cannot be categorised as 'private' or 'professional' use and almost 36% where the possibility of 'personal' usage is much higher. If we combine this with the other categories, more than two thirds of the usage data can be explained!

Table 4 Usage data of all books, refined

Usage	Total	National	International
Academic	27.82%	4.71%	23.11%
Non-profit	0.91%	0.19%	0.72%
Government	2.25%	0.30%	1.95%
Business	1.18%	0.20%	0.98%
ISP	31.86%	0.75%	31.11%
ISP (High internet usage)	35.97%	8.68%	27.29%
Total		*14.84%*	*85.16%*

Figure 4 Downloads OAPEN Library - 2011; refined



7.7 Possible influences on usage

In the next paragraphs, two possible influences on the usage in the OAPEN Library will be discussed: subject and language. Using the average number of downloads per group of titles, the distribution of the providers will be analysed. The average number of downloads is used here to compensate for the varying number of titles per subject or language. As described below, the number of titles with the same subject ranges from 65 to 22 titles. The same holds true for titles in the same language: the set contains 460 books in English; 105 in Dutch; 112 in Italian and 126 written in German.

After analysing the data on the level of the complete OAPEN Library or relative large subsets, the data on the level of individual books will be discussed. However, the analysis at the individual level will be less thorough.

7.7.1 Subject – highest level

In the OAPEN Library, the subject of the books is described using the BIC classification (Book Industry Communication, 2010). Due to its hierarchical nature, the classification assigned to each book can be abbreviated. This results in a larger group of monographs which share the same – broad – subject. The usage data of the 10 largest groups were compared with the averages of all books in the OAPEN Library, to see if the usage patterns differ significantly. In the following table, all data is normalised to the average number of downloads per subject.

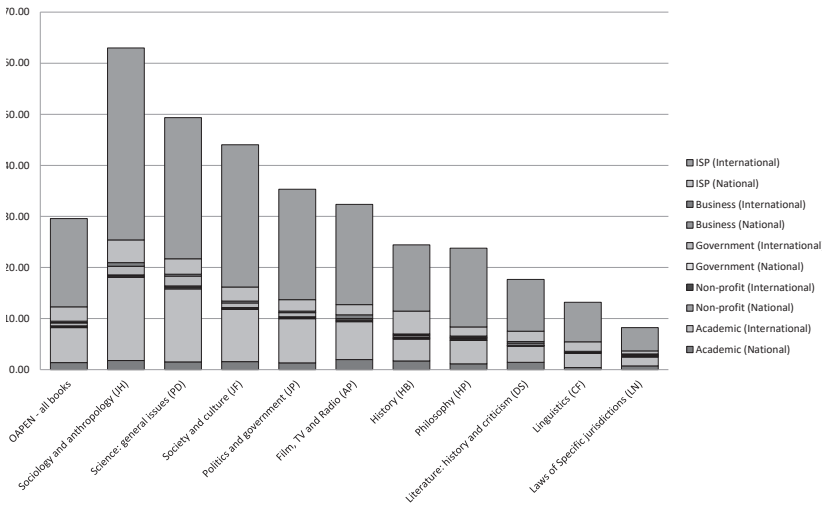
Table 5 Subject: usage data of 10 largest groups

Number of titles	Book Subject	Academic (National)	Academic (International)	Non-profit (National)	Non-profit (International)	Government (National)	Government (International)	Business (National)	Business (International)	ISP (National)	ISP High Internet usage (National)	ISP (International)	ISP High Internet usage (International)	Total
859	OAPEN - all books	1.39	6.84	0.06	0.21	0.09	0.58	0.06	0.29	0.22	2.57	9.20	8.07	29.58
65	Sociology and anthropology (JH)	1.78	16.35	0.03	0.29	0.11	1.68	0.02	0.69	0.11	4.35	19.35	18.22	62.98
43	Science: general issues (PD)	1.51	14.30	0.00	0.37	0.21	1.91	0.00	0.40	0.00	3.02	13.74	13.88	49.35
51	Society and culture (JF)	1.57	10.27	0.04	0.24	0.06	0.88	0.00	0.37	0.14	2.63	15.08	12.75	44.02
148	Politics and government (JP)	1.32	8.67	0.03	0.22	0.12	0.80	0.01	0.28	0.11	2.14	12.66	8.97	35.32
30	Film, TV and Radio (AP)	1.97	7.43	0.00	0.20	0.00	0.33	0.07	0.73	0.00	2.00	10.50	9.13	32.37
151	History (HB)	1.72	4.27	0.09	0.15	0.17	0.34	0.11	0.16	0.44	4.03	6.97	5.99	24.42
22	Philosophy (HP)	1.14	4.59	0.09	0.23	0.00	0.18	0.14	0.23	0.27	1.50	6.82	8.59	23.77
28	Literature: history and criticism (DS)	1.43	3.18	0.07	0.07	0.00	0.04	0.39	0.32	0.00	2.04	3.86	6.29	17.68
22	Linguistics (CF)	0.41	2.82	0.00	0.09	0.05	0.18	0.00	0.05	0.50	1.36	3.82	3.91	13.18
22	Laws of Specific jurisdictions (LN)	0.73	1.73	0.14	0.18	0.05	0.14	0.00	0.14	0.09	0.50	2.64	1.91	8.23

7.7.1.1 Average downloads per subject

All data is normalised to the average number of downloads per subject.

Figure 5 Average downloads per subject



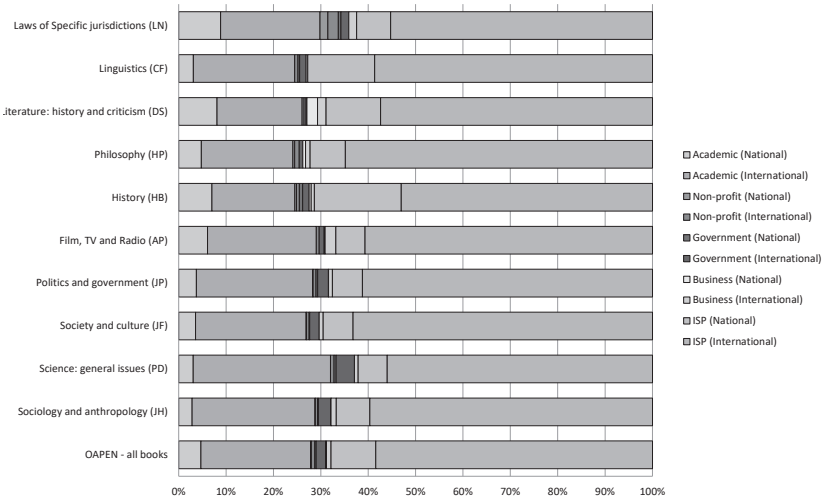
When looking at the average number of downloads, it is striking that subjects from the social sciences – Sociology and anthropology, Society and culture, Politics and government – are more 'popular' than well-known subjects from the humanities, such as History, Philosophy and Literature.

The large differences in downloads per subject raise the question whether this is caused by differences in the usage per readers group. For instance, is the large uptake on Sociology and anthropology caused by relative high academic usage? In order find the answer, the percentages of usage per provider was computed.

7.7.1.2 Average downloads per subject – percentage

All data is normalised to the average number of downloads per subject.

Figure 6 Average downloads per subject - percentage



Here, the distribution across the subjects does not change dramatically, with the exception of History, Linguistics and Literature. For these subjects the academic *national* usage is relatively high. In the case of History, the explanation may lie in the fact that if the historic subject is national, the usage will tend to be national as well. Linguistics and Literature are of course closely bound to national languages; the percentage of academic readers interested in their national language will be greater than readers interested in foreign languages.

Furthermore, the largest percentages of national 'ISP usage' coming from countries with a high number of internet usage – in other words: readers that are most likely to be interested for non-professional reasons – are to be found with History and Linguistics. In contrast, the usage of legal books (Laws of Specific jurisdictions) by government agencies and businesses is relatively high, but is still dwarfed by academic and 'ISP usage'.

7.7.2 Language – highest level

The collection of the OAPEN Library contains several languages. Not all languages are equally represented. Therefore, only the largest groups are discussed. In the following table, all data is normalised to the average number of downloads per language.

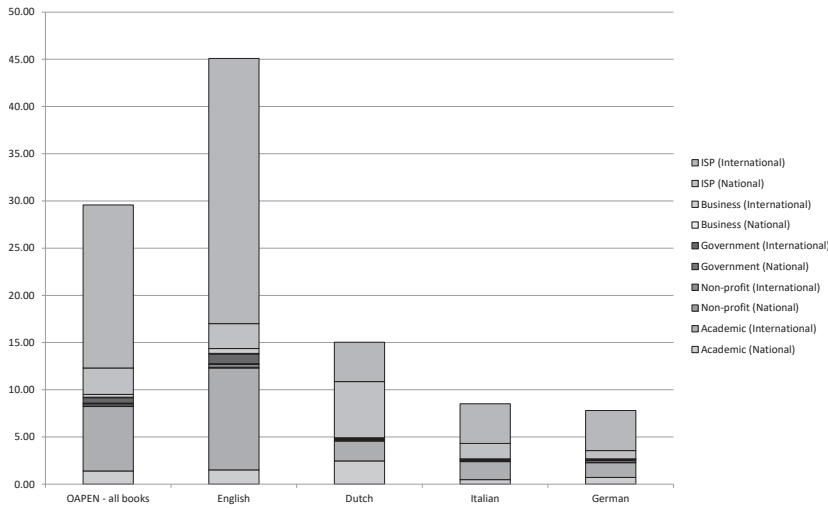
Table 6 Language: usage data of 4 largest groups

Number of titles	Language	Academic (National)	Academic (International)	Non-profit (National)	Non-profit (International)	Government (National)	Government (International)	Business (National)	Business (International)	ISP (National)	ISP High internet usage (National)	ISP (International)	ISP High internet usage (International)	Total
859	OAPEN - all books	1.39	6.84	0.06	0.21	0.09	0.58	0.06	0.29	0.22	2.57	9.20	8.07	29.58
460	English	1.51	10.80	0.07	0.32	0.07	1.00	0.10	0.51	0.02	2.63	15.38	12.70	45.09
105	Dutch	2.46	2.10	0.08	0.01	0.08	0.10	0.03	0.06	0.00	5.94	1.62	2.56	15.04
112	Italian	0.48	1.90	0.00	0.07	0.07	0.11	0.01	0.04	1.64	0.00	1.97	2.22	8.52
126	German	0.71	1.56	0.06	0.17	0.08	0.06	0.02	0.02	0.00	0.88	1.54	2.69	7.79

7.7.2.1 Average downloads per language

All data is normalised to the average number of downloads per language.

Figure 7 Average downloads per language



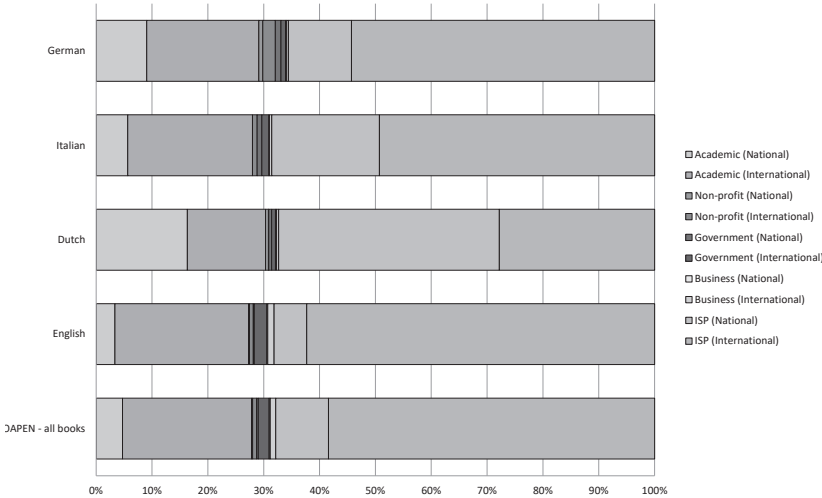
When looking at the total number of downloads per average title, it becomes clear that English is the most read language: it amounts to approximately 150% of the average of the complete OAPEN Library usage. The average downloads of Dutch titles is almost twice as high as the number of downloads for titles in Italian and German. The explanation may be found in the fact that 14% of all usage data originated in the Netherlands, while Italian providers are responsible for 4% and German providers for 8%.

The differences in usage may also be connected to differences in usage by each reader group. For that reason, the percentages of usage per provider was computed.

7.7.2.2 Average downloads per language – percentage

All data is normalised to the average number of downloads per language.

Figure 8 Average downloads per language - percentage



The percentages reveal the 'national appeal' of Dutch language titles: the percentages of national usage – both academic and coming from ISPs – are far greater than the other languages, or the average of all books. The international usage coming from ISPs is by far the lowest, and the percentage of international academic use is also lower compared to the other languages. National usage for Dutch language books is of course coming from both the Netherlands and from Belgium. In this particular case, the Dutch language books published by *Dutch* publishers account for 35% of the usage data, while the Dutch language books published by *Belgian* publishers account for 4.5% of the usage.

In contrast, the books written in English have the lowest percentages of national usage. This is of course not surprising: English functions as the 'lingua franca' of science. The percentages of German and Italian books fall between these two extremes. From this we might conclude that books written in English, German and Italian appeal to a far more international audience than those written in Dutch. If Dutch or Belgian authors want their work to be used outside their countries, translation is necessary. The same effect was found for Danish, but the number of titles was much lower: 22. Therefore these titles were not taken into account here.

7.7.3 Subject – book level

Here, all downloads per individual title are analysed, per subject. The main goal is to look at the skewedness of the total number of downloads: is it heavily influenced by just a few titles, or is the number of downloads spread relatively even? Furthermore, the usage percentages of the 15 most downloaded titles are visualised, in order to determine if they deviate greatly from the percentages of the whole group.

It becomes clear that the social sciences are more prone to skewed distributions of downloads, compared to humanities. The groups Sociology and anthropology, Society and culture and Politics and government all contain a title that is downloaded far more than the rest. All these titles with an exceptional number of downloads were authored by members of IMISCOE Research Network.[3] The website of the IMISCOE Network contains links to all books in the OAPEN Library. This may be the reason for the high number of downloads.

When we look at the usage percentages, we see that lower number of downloads seem to correlate with higher differences in percentages. A good example can be found in the group Laws of Specific jurisdictions, where the title *Videovernehmung kindlicher Zeugen ; zur Praxis des Zeugenschutzgesetzes,* ISBN 9783938616833 shows a usage percentage of 40% by foreign government organisations. This looks very spectacular, but it is caused by 2 downloads. Small differences give high percentages!

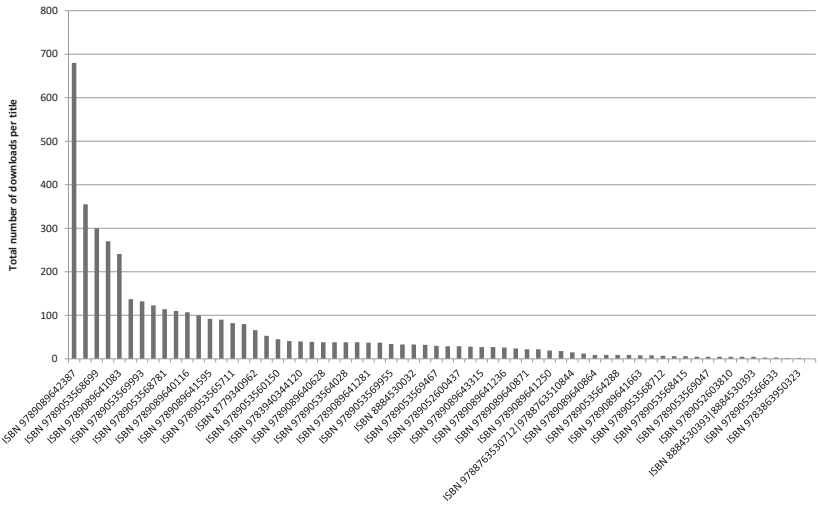
Each book is identified using ISBN (International Standard Book Number).

7.7.3.1 Sociology and anthropology

The chart depicts the total number of downloads per title. The total number of titles is 65.

3 See: http://www.imiscoe.org

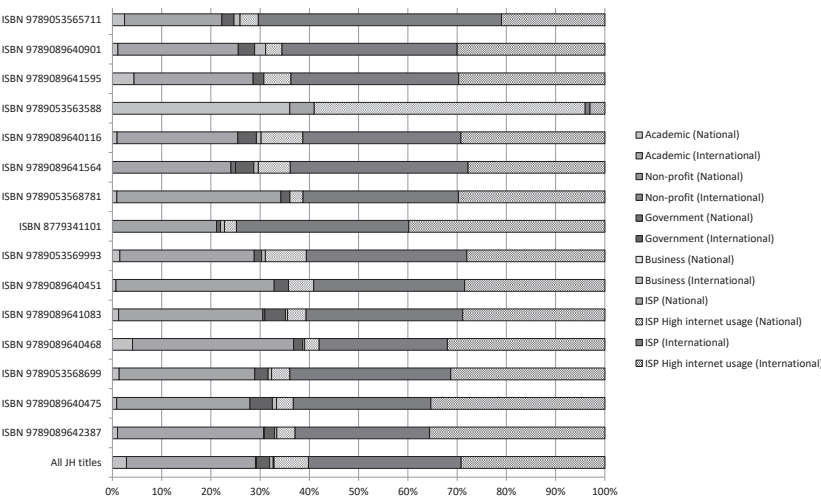
Figure 9 Sociology and anthropology (JH) - Total downloads per title



Here we see an outlier, with 680 downloads: *Diaspora and Transnationalism : Concepts, Theories and Methods*, ISBN 9789089642387.

The chart depicts the 15 most downloaded titles, combined with the percentages for all titles with the same subject.

Figure 10 Sociology and anthropology (JH) - Most downloaded, percentage



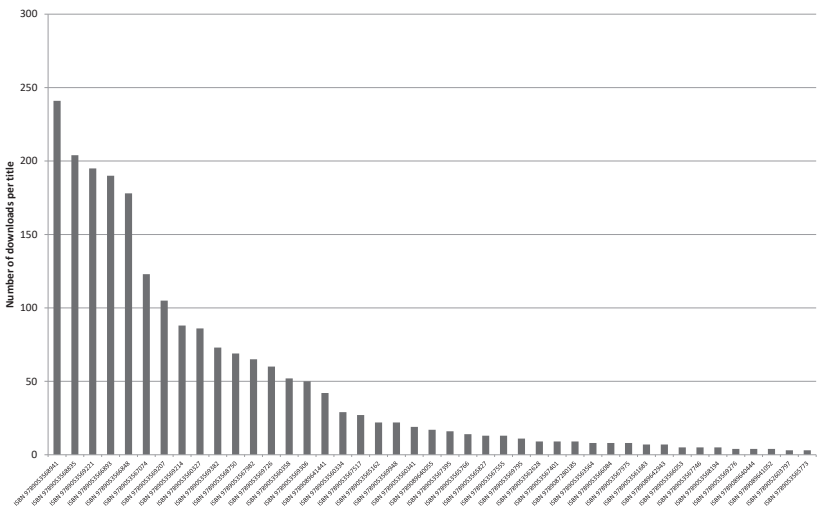
The first 15 titles are responsible for 71.64% of all downloads. Here we see one obvious outlier: *Nationale identiteit en meervoudig verleden,* ISBN

9789053563588. The difference may come from the large amount of OAPEN users from the Netherlands and Belgium.

7.7.3.2 Science: general issues

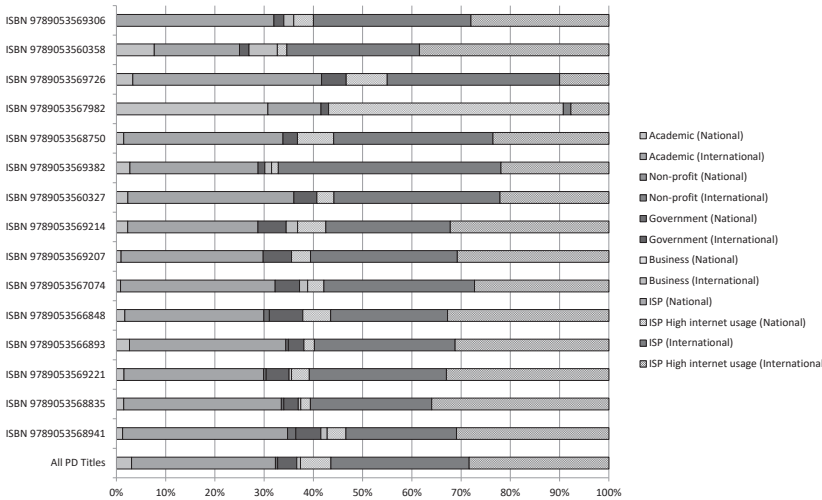
The chart depicts the total number of downloads per title. The total number of titles is 43.

Figure 11 Average downloads per subject - percentage



The chart depicts the 15 most downloaded titles, combined with the percentages for all titles with the same subject.

Figure 12 Science: general issues (PD) - Most downloaded, percentage

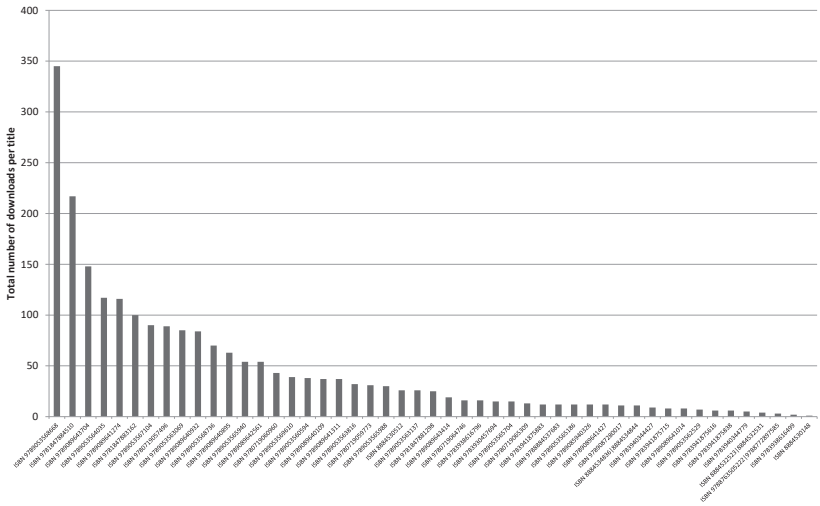


The first 15 titles are responsible for 83.84% of all downloads. Here we see the same pattern as the previous subject: no large differences save one outlier: *Van natuurlandschap tot risicomaatschappij : De geografie van de relatie tussen mens en milieu*, ISBN 9789053567982. As this is the only Dutch language title, the large amount of Dutch OAPEN users may have caused this.

7.7.3.3 Society and culture

The chart depicts the total number of downloads per title. The total number of titles is 51.

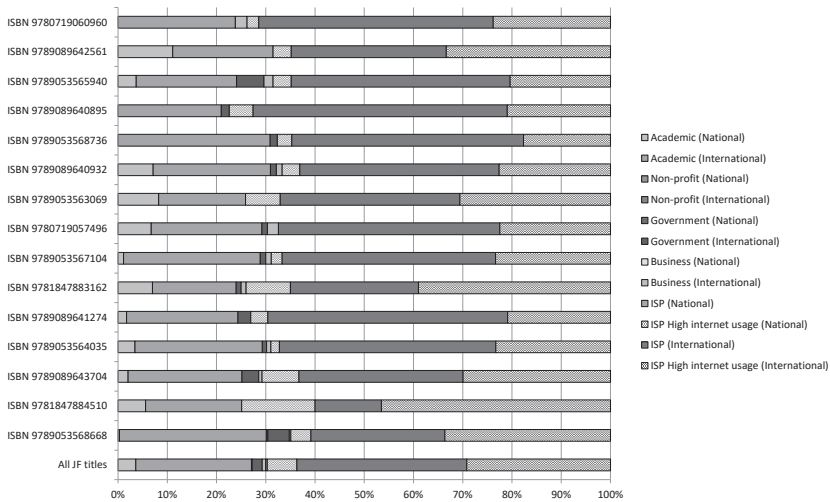
Figure 13 Society and culture (JF) - Total downloads per title



This group contains one outlier: *The Dynamics of International Migration and Settlement in Europe : A State of the Art*, ISBN 9789053568668.

The chart depicts the 15 most downloaded titles, combined with the percentages for all titles with the same subject.

Figure 14 Society and culture (JF) - Most downloaded, percentage

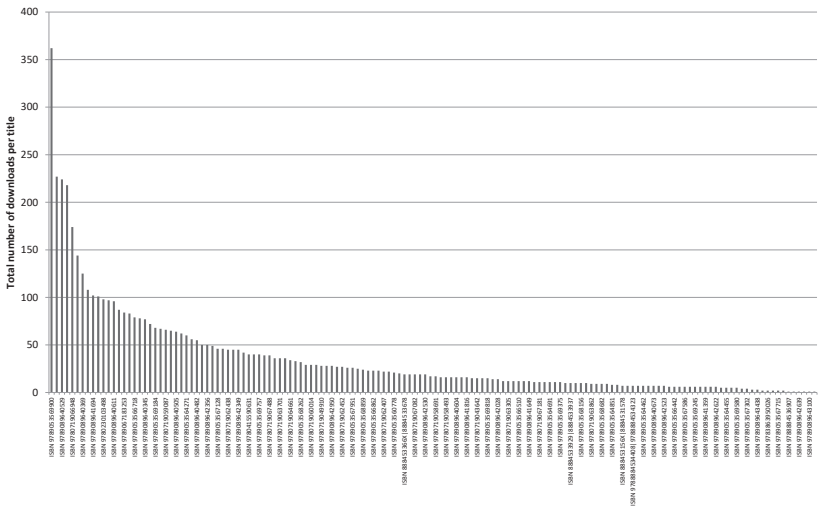


The first 15 titles are responsible for 75.08% of all downloads. There is no obvious outlier.

7.7.3.4 Politics and government

The chart depicts the total number of downloads per title. The total number of titles is 148.

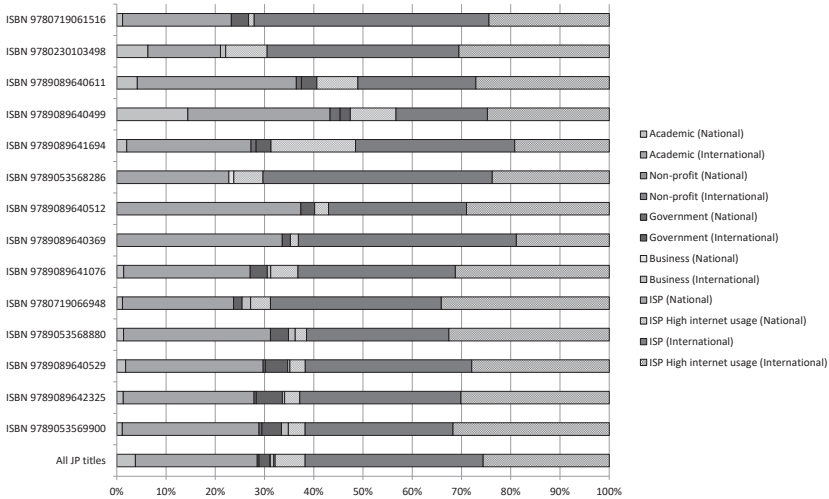
Figure 15 Politics and government (JP) - Total downloads per title



As is the case with other social science groups, here we see one outlier: *Innovative Concepts for Alternative Migration Policies : Ten Innovative Approaches to the Challenges of Migration in the 21st Century*, ISBN 9789053569900

The chart depicts the 15 most downloaded titles, combined with the percentages for all titles with the same subject.

Figure 16 Politics and government (JP) - Most downloaded, percentage

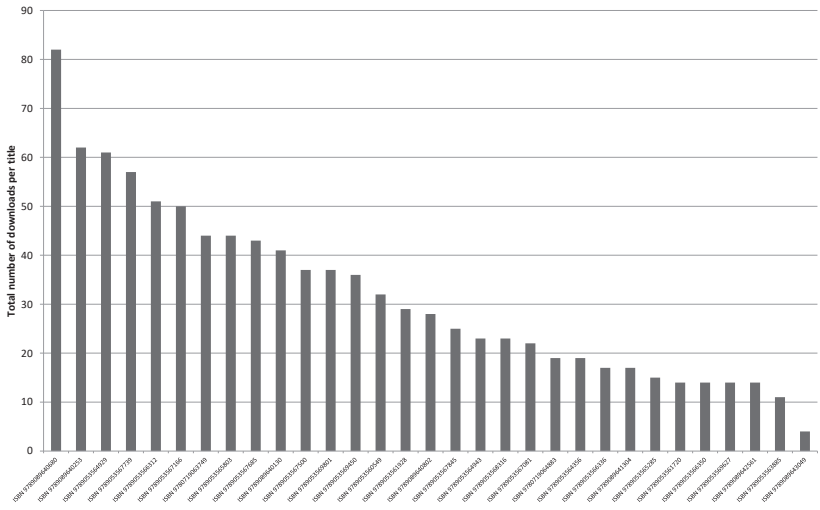


The first 15 titles are responsible for 42.98% of all downloads. The title *Illegal Residence and Public Safety in the Netherlands*, ISBN 9789089640499 has a relative large percentage of national academic usage, which is not surprising given the fact that is was published in the Netherlands.

7.7.3.5 Film, TV and Radio

The chart depicts the total number of downloads per title. The total number of titles is 30.

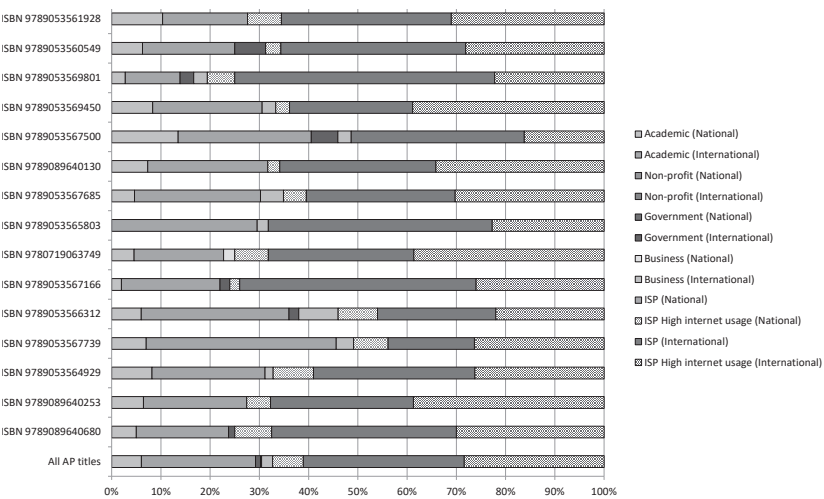
Figure 17 Film, TV and Radio (AP) - Total downloads per title



There is no obvious outlier.

The chart depicts the 15 most downloaded titles, combined with the percentages for all titles with the same subject.

Figure 18 Film, TV and Radio (AP) - Most downloaded, percentage



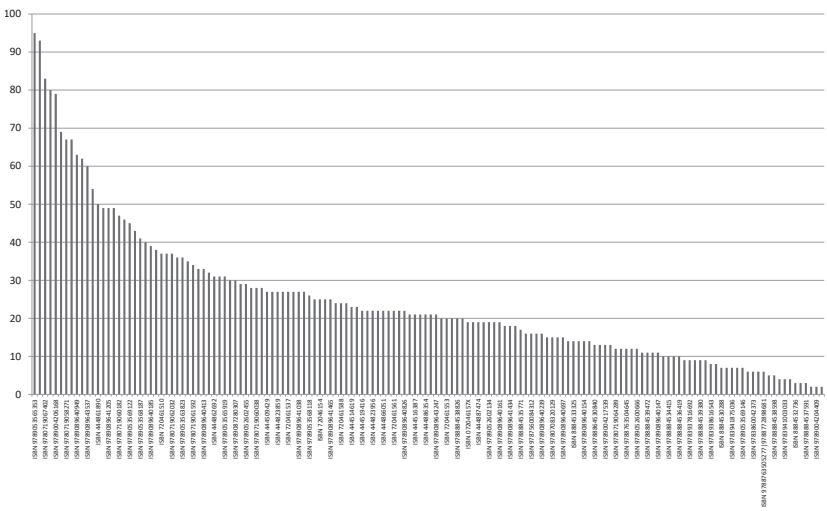
The first 15 titles are responsible for 71.68% of all downloads. Here, downloads from government agencies are a relatively large percentage

of all downloads for one title. This is caused by the overall low number of downloads: per title it is one or two downloads.

7.7.3.6 History

The chart depicts the total number of downloads per title. The total number of titles is 151.

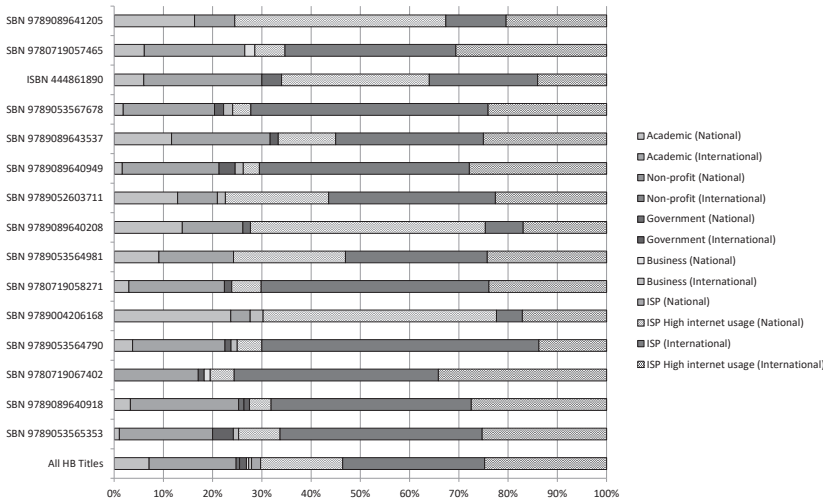
Figure 19 History (HB) - Total downloads per title



There is no obvious outlier.

The chart depicts the 15 most downloaded titles, combined with the percentages for all titles with the same subject.

Figure 20 History (HB) - Most downloaded, percentage



The first 15 titles are responsible for 27.66% of all downloads. We can see a relative high number of national high internet downloads for these titles:

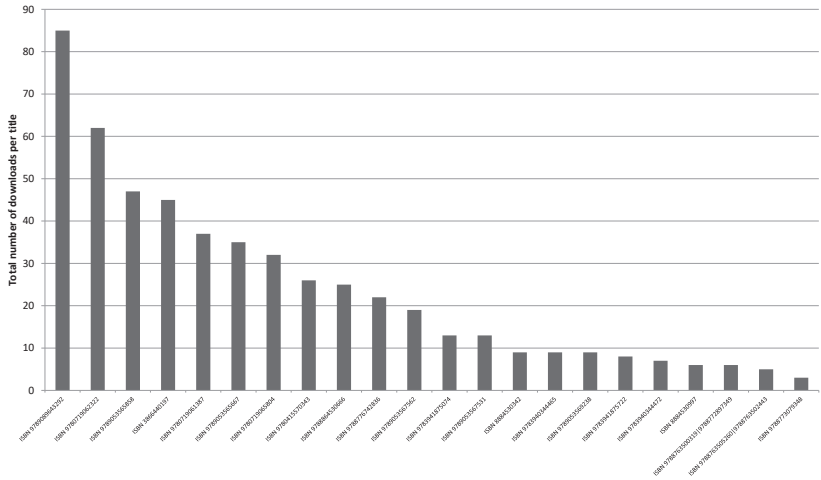
- *Literary Cultures and Public Opinion in the Low Countries, 1450-1650*, ISBN 9789004206168
- *De hand van Huizinga*, ISBN 9789089640208
- *Het Hemels Mandaat : De Geschiedenis van het Chinese Keizerrijk*, ISBN 9789089641205
- *Opera omnia Desiderii Erasmi : Ordinis secundi tomus quartus*, ISBN 444861890

All are published by Dutch publishers, and the language is either Dutch or the book is concerned with a Dutch subject.

7.7.3.7 Philosophy

The chart depicts the total number of downloads per title. The total number of titles is 22.

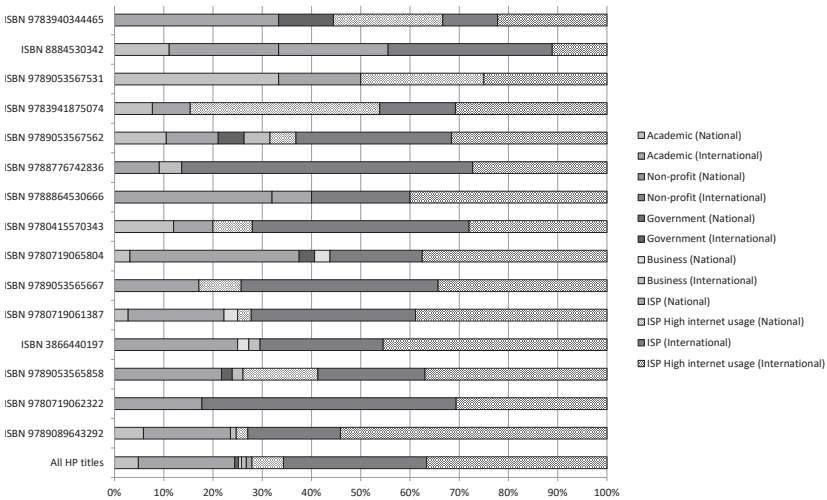
Figure 21 Philosophy (HP) - Total downloads per title



There is no obvious outlier.

The chart depicts the 15 most downloaded titles, combined with the percentages for all titles with the same subject.

Figure 22 Philosophy (HP) - Most downloaded, percentage

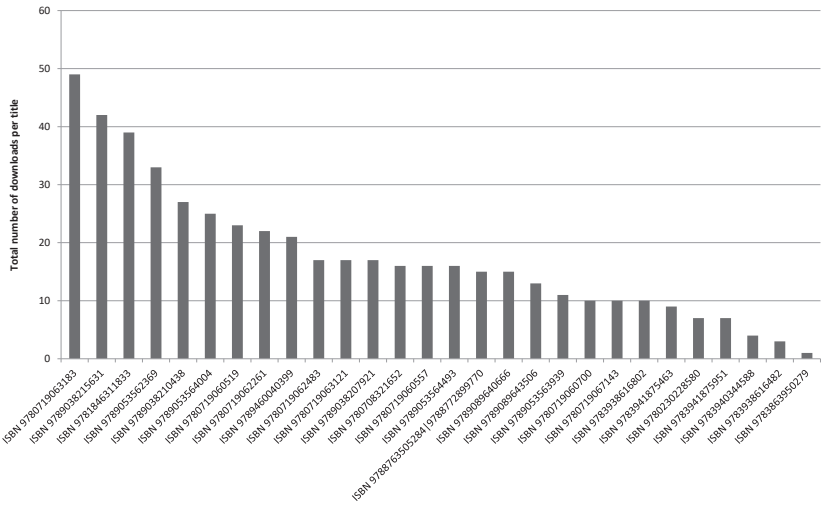


The first 15 titles are responsible for 91.59% of all downloads. The 15th title received 9 downloads. From now, we will see strong deviations in the percentages, combines with a small number of down loads

7.7.3.8 Literature: history and criticism

The chart depicts the total number of downloads per title. The total number of titles is 28.

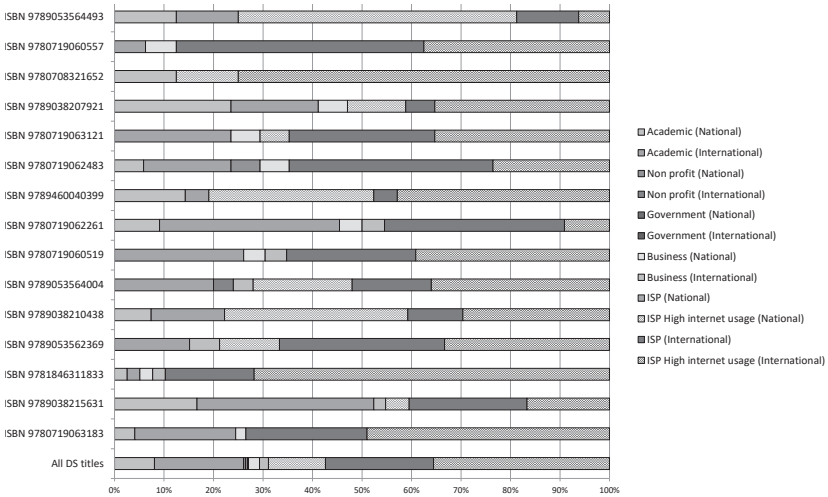
Figure 23 Literature: history and criticism (DS) - Total downloads per title



There is no obvious outlier.

The chart depicts the 15 most downloaded titles, combined with the percentages for all titles with the same subject.

Figure 24 Literature: history and criticism (DS) - Most downloaded, percentage

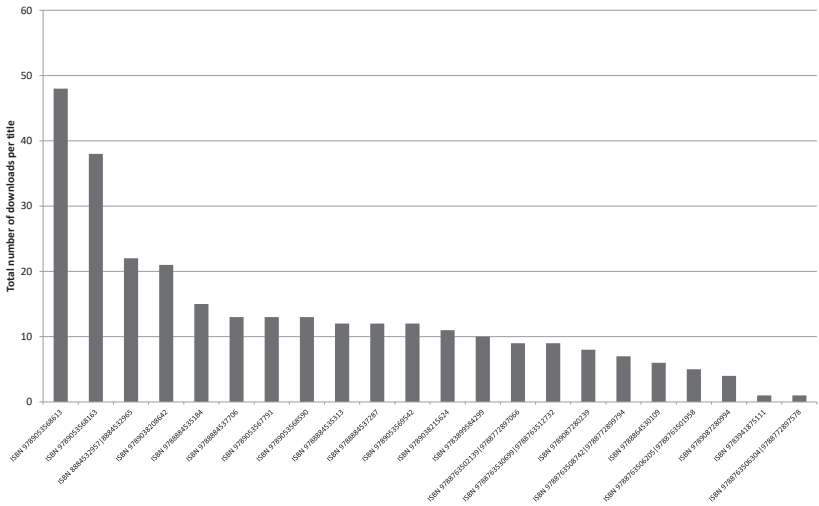


The first 15 titles are responsible for 76.77% of all downloads. Here we see much variation in the usage percentages per title. Because of the relative low number of downloads – ranging from 49 to 16 – one download has a large impact in the chart.

7.7.3.9 Linguistics

The chart depicts the total number of downloads per title. The total number of titles is 22.

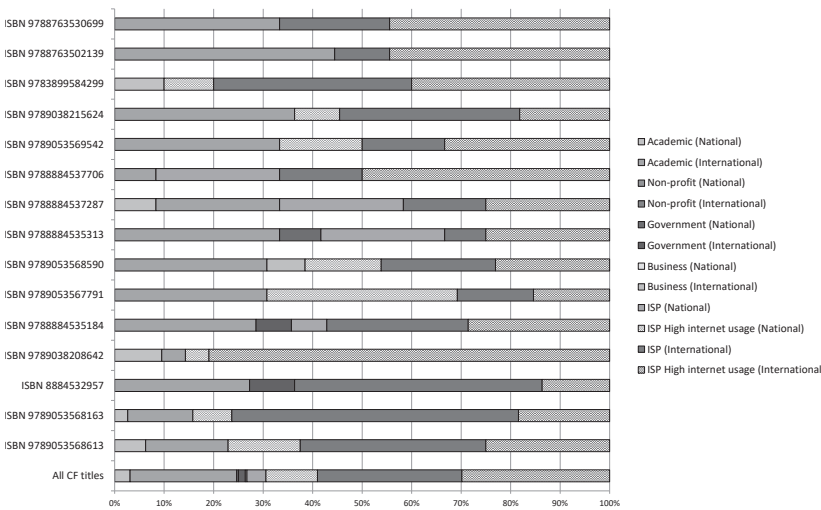
Figure 25 Linguistics (CF) - Total downloads per title



The first two titles are responsible for 29.66% of all downloads.

The chart depicts the 15 most downloaded titles, combined with the percentages for all titles with the same subject.

Figure 26 Linguistics (CF) - Most downloaded, percentage

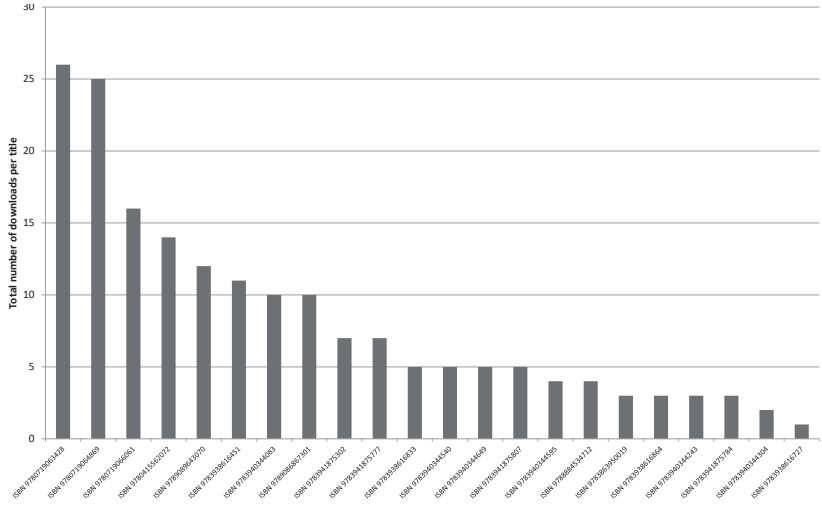


The first 15 titles are responsible for 88.97% of all downloads. Again, we see a large difference in usage percentages, but a small number of overall downloads.

7.7.3.10 Laws of Specific jurisdictions

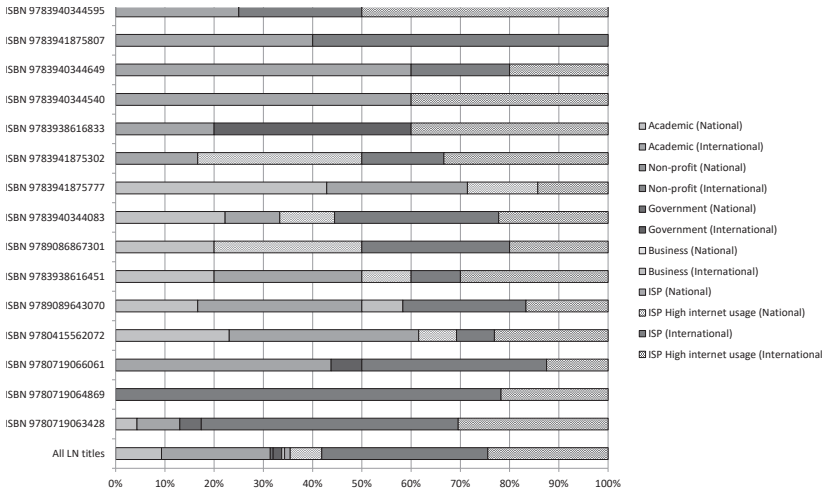
The chart depicts the total number of downloads per title. The total number of titles is 22.

Figure 27 Laws of Specific jurisdictions (LN) - Total downloads per title



The chart depicts the 15 most downloaded titles, combined with the percentages for all titles with the same subject.

Figure 28 Laws of Specific jurisdictions (LN) - Most downloaded, percentage



The first 15 titles are responsible for 89.50% of all downloads. Again, we see a large difference in usage percentages, but a small number of overall downloads.

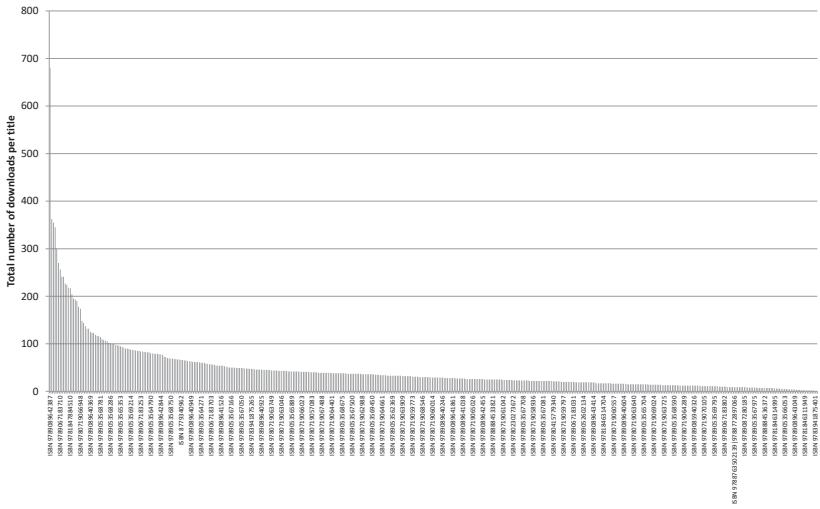
7.7.4 Language – book level

The analysis on languages shows the same pattern: a low number of downloads seems to be correlated with high diversity in percentages. This is best illustrated with the differences between English and German. The usage percentages for English – where the average number of downloads per book is 45.29 – are not much different. This contrast with German, where the average number of downloads is much lower: 7.79.

7.7.4.1 English

The chart depicts the total number of downloads per language. The total number of titles is 460.

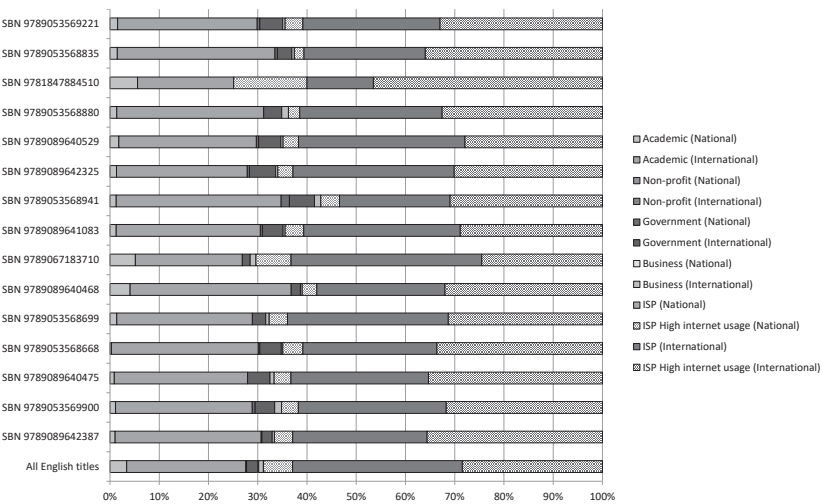
Figure 29 English - Total downloads per title



The outlier is of course: *Diaspora and Transnationalism : Concepts, Theories and Methods*, ISBN 9789089642387.

The chart depicts the 15 most downloaded titles, combined with the percentages for all titles with the same language.

Figure 30 English - Most downloaded, percentage

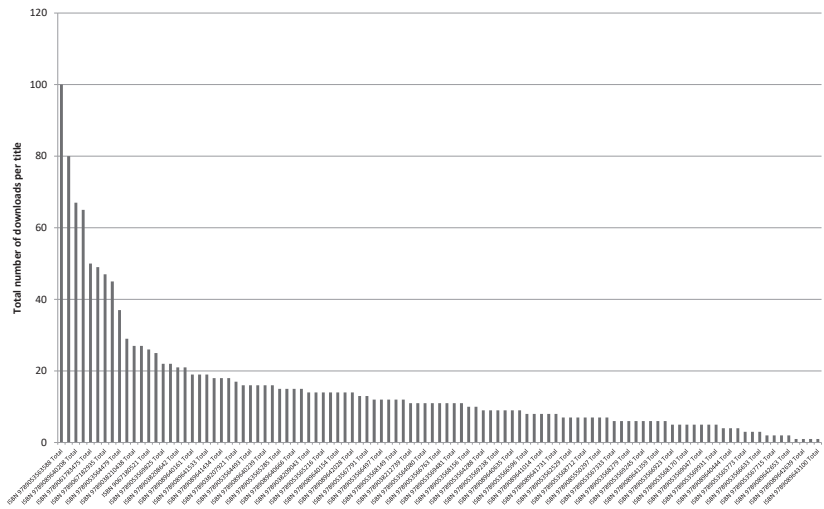


The first 15 titles are responsible for 20.90% of all downloads. Here the usage percentages are the most consistent.

7.7.4.2 Dutch

The chart depicts the total number of downloads per language.

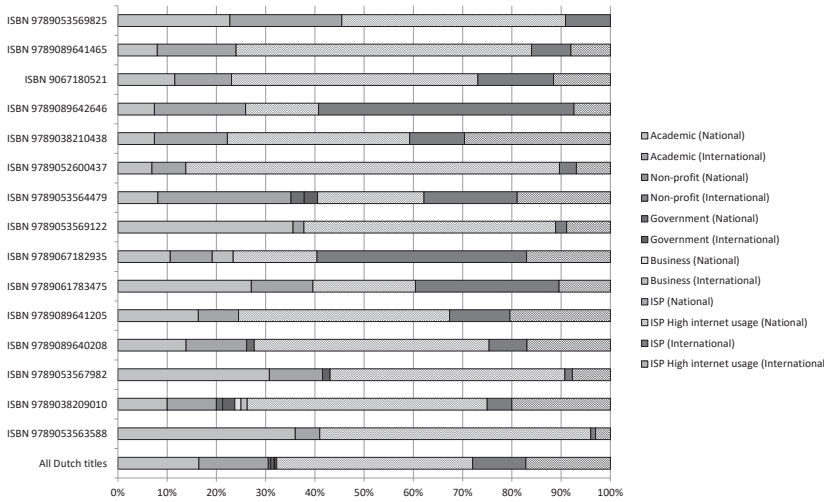
Figure 31 Dutch - Total downloads per title



The most popular title is *Nationale identiteit en meervoudig verleden,* ISBN 9789053563588

The chart depicts the 15 most downloaded titles, combined with the percentages for all titles with the same language.

Figure 32 Dutch - Most dowloaded, percentage

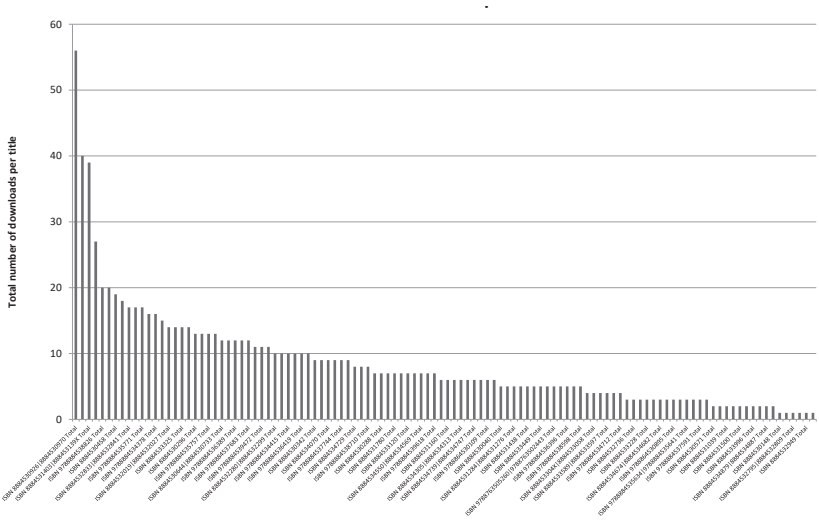


The first 15 titles are responsible for 44.08% of all downloads. The 'national' appeal – which was discussed before – is clearly visible through the relative high percentages of national academic and ISP usage.

7.7.4.3 Italian

The chart depicts the total number of downloads per language. The total number of titles is 112.

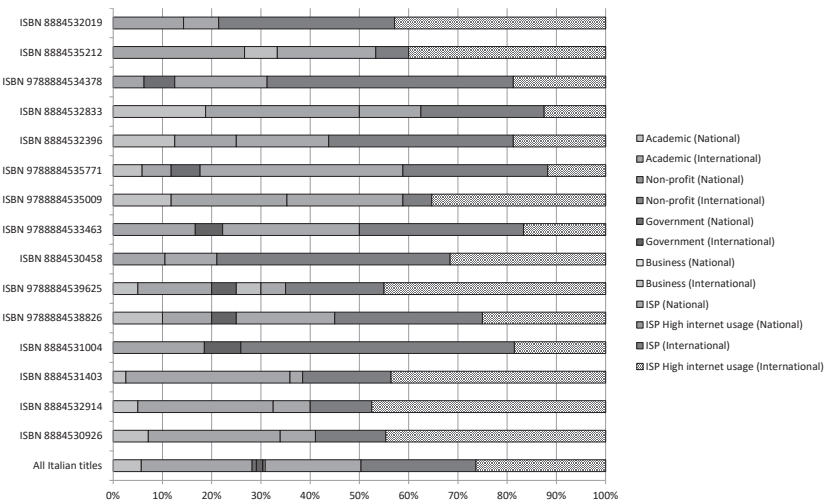
Figure 33 Italian - Total downloads per title



The most downloaded title is *Tell Barri/Kahat: la campagna del 2000*, ISBN 8884530926.

The chart depicts the 15 most downloaded titles, combined with the percentages for all titles with the same language.

Figure 34 Italian - Most downloaded, percentage

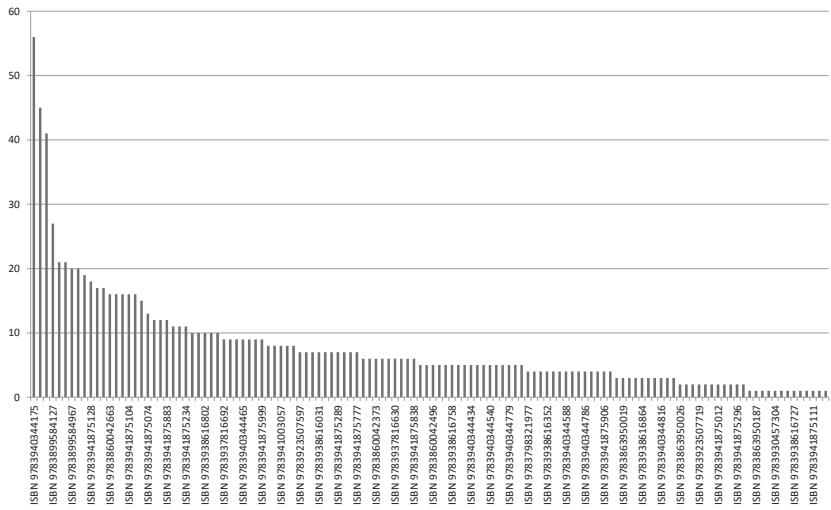


The first 15 titles are responsible for 36.79% of all downloads. The small number of downloads correlates once again with larger differences in usage percentages.

7.7.4.4 German

The chart depicts the total number of downloads per language. The total number of titles is 126.

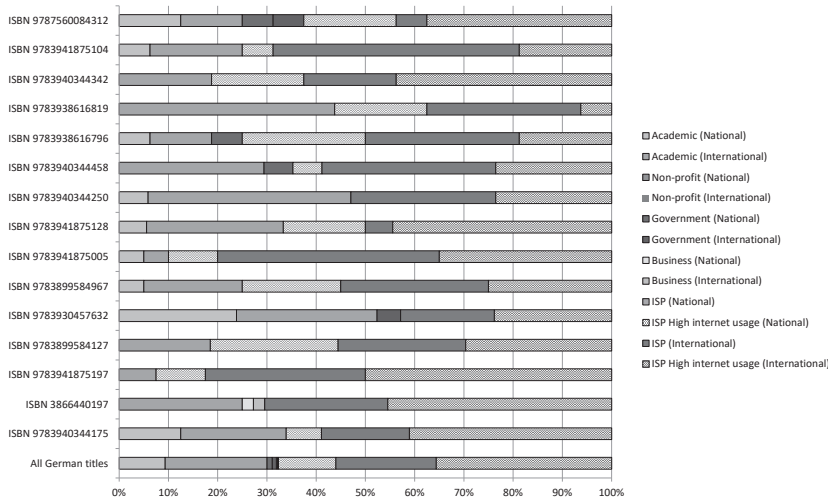
Figure 35 German - Total downloads per title



The most popular title – with 56 downloads – is *Ein Compendium sumerisch-akkadischer Beschwörungen*, ISBN 9783940344175.

The chart depicts the 15 most downloaded titles, combined with the percentages for all titles with the same language.

Figure 36 German - Most downloaded, percentage



The first 15 titles are responsible for 37.68% of all downloads.

7.8 Conclusion

7.8.1 The method as addition to existing assessments

The problem addressed in this article is the measurement of scientific impact and societal relevance in the field of Humanities and Social Sciences. When looking at methods to measure scientific impact through the published output, we saw that the standard bibliometric methods employed in the field of Science, Technology and Medicine – where publishing in articles is the norm – do not function well in a field where monographs are the standard. Quantitative methods that do take monographs into account are aimed at measuring scientific impact only, leaving out the societal benefits.

The methods used to assess societal relevance also have drawbacks. First of all, most of these methods are qualitative, depending on self-assessments by scholars and on opinions by representatives of stakeholders outside academia. Apart from possible subjective biases, these methods require that stakeholders are known. This is not always the case in the field of HSS, especially in the humanities. Another aspect is the amount of labour involved: discussions with focus groups, sending questionnaires and conducting desk

research and stakeholder analysis requires quite a lot of manpower. Societal relevance is hard to measure in the field of HSS, especially when the groups that would primarily benefit from research are not always known. In the field of STM, patents are used as an indicator, but a comparable indicator for HSS research has not been defined.

In this article, a new method is used to overcome some of those issues. This method measures the usage of monographs and identifies the organisation responsible for the internet access. Therefore, both the usage and the readers of each monograph are known. The amount of usage – here restricted to number of downloads – by each type of reader could be used to assess the value of scientific or scholarly output.

The method is quantitative, which makes the results easier to validate. The amount of measurements is also large: the data set for this article consists of over 25,000 downloads by more than 1,500 providers, spread over 859 monographs. A large data set reduces the chances of outliers influencing the results. It is not necessary to know the stakeholders in advance: the method is used to identify the readers. This solves one of the identified problems: especially in the field of humanities, where benefactors besides academics are not always known. Knowing other users besides academics makes it easier to assess the societal relevance. Another drawback of the described qualitative methods is the labour intensity; by relying heavily on automated tools, this method is relatively easy to execute. Furthermore, one of the problems attached to measuring societal relevance is attribution: how to measure the influence of a certain scholar? Here we look at the usage of books, which makes it easy to identify the influence of each author.

7.8.2 Discussion of the results

When looking at the results, it becomes clear that the monographs are not used exclusively by scholars. From the measured data, over 27% is directly linked to academic users. The percentage of usage that can be linked directly to other 'professional' users is quite small: less than 5%. This leaves a large portion of users that cannot be categorised immediately. By taking into account the percentage of ISPs per country, this group is further categorised. This results in a group of users that cannot be categorised and a group of users – more than 35% of all users – that have a higher probability to be 'non-professional users', also known as the 'general public'. Taken together, more than 68% of the usage can be categorised, and almost 45% of all usage comes from non-academics. This might indicate that the monographs have an impact in society.

In order to further refine the results, two possible influences on monograph usage were analysed: subject and language. When looking at the influence of subject on usage, we saw that the average number of downloads per subject varies widely. Most of the subjects that received a higher number of downloads than the average of the total set come from the field of the social sciences. The humanities were less 'popular', with amounts that lie mostly below the average for the complete set. If we use this as a measure of societal relevance, we might conclude that monographs in the social sciences enjoy a relatively large readership outside academia. The number of books on a certain subject may have influenced these results, but it is not very likely: 65 books on Sociology and anthropology receive an average number of 62.98 downloads, and 51 books on Society and culture are downloaded 44.02 times on average. In contrast, 151 History books – a much larger amount of titles – are downloaded 24.42 times on average. If the usage percentages per group were taken into account, it becomes clear that they do not differ significantly. Only History, Linguistics and Literature are the exception: here the percentage of 'national' usage is higher. These subjects might have a tendency to be bound to national borders.

In order to measure the influence of language on monograph usage, the four largest language groups were analysed. Again, the average number of downloads and the percentages per groups were used. It was hardly surprising to discover that books in English – the 'lingua franca' of science – were downloaded the most. A more interesting discovery was the fact that some languages such as Dutch (and Danish) were read much less outside of national borders that Italian or German. While a Dutch or Belgian scholar would need a translation in order to have more influence abroad, this does apply far less for Germans or Italians.

The analysis on the level of individual books revealed that within the social sciences, the distribution of usage was relatively more skewed than within the humanities: the groups Sociology and anthropology, Society and culture and Politics and government all contain a title that is downloaded far more than the rest. It is interesting to note that all these books were written by authors connected to the IMISCOE[4] network. Possibly, readers were alerted through the IMISCOE website.

The results give an indication of the usage, and it becomes clear that HSS monographs are read outside academia, proving the societal relevance. Below, the conclusions are discussed a little further.

4 See: http://www.imiscoe.org

First of all, the research method is based on measuring usage of electronic versions of monographs. The usage data of the paper versions – such as sales figures or borrowing data from libraries – were not available. It would be interesting to see if the percentage of user categories would differ dramatically. Given the economic circumstances discussed in the first paragraph, we might conclude that the dissemination of paper books is far less successful that electronic ones. However, it may be possible that a certain group of readers prefers the paper monograph to the electronic version, and this aspect has not been taken into account.

Another aspect is the dissemination channel. In earlier research done by the author (Snijder, 2010), it became clear that different dissemination channels display different results. There, the usage through an institutional repository was significantly smaller than usage through the Google Book Search program. Here, one dissemination channel is used and therefore we cannot compare the usage patterns. In other words: we cannot determine if the low usage by government agencies, non-profit organisations and businesses is solely caused by the contents of the monographs, or whether it is partly caused by the fact that the OAPEN Library is not used by these types of organisations. Another aspect of the OAPEN Library is that it only hosts open access monographs. This means that the complete text of the books is fully available online. As there is no comparable data set available of monographs that are not fully accessible, we cannot determine how usage is influenced by open access.

The data analysed is the usage measured through the OAPEN website; direct downloads are not taken into account. At this point, only the total number of downloads is available – no other data. When the complete data becomes available, it will be interesting to see whether the percentage of 'ISP' usage will become smaller. The total number of downloads in 2011 – over 300,000 – is more than 6 times higher than the number of downloads in the current data set. This much larger number of 'direct' downloads may come from library systems or other collections of book data. These data files will probably have been made available to 'professional' users, such as academics or civil servants. This may explain the small percentage of government use, or it may uncover a much higher scientific and scholarly use.

7.8.3 Possible refinements to the method

The method in its current form uses relatively broad categories. Users are divided into 6 groups and are categorised as national or international. Based

on these categories, it is simple to make analyses on an abstract level. By doing so, smaller effects based on specific books are not visible. One of the possible refinements could be categorisation on countries. This enables us to look at a more detailed level. For instance, the books published by KITLV Press are mostly downloaded through Indonesian providers. The reason for this is clear: the subject of all KITLV titles is South East Asia, and most of those monographs describe themes from Indonesia. An analysis on country level may uncover more of these effects, but the level of detail required is beyond the scope of this article.

Another refinement could be found in analysing the usage patterns for each individual author. Before, the usage per subject has been analysed. We could use the percentages of the groups of readers as a 'baseline' to compare the usage patterns of the work or works from a certain author. Again, we cannot be sure how the dissemination channel influences the results. Therefore, this kind of analysis should be done with caution, and preferably at a time where more experience with using this method has been gained.

The data set of this article is available at:
http://www.persistent-identifier.nl/?identifier=urn:nbn:nl:ui:13-fbfa-yd.

7.8.4 Evaluation of the results

In this article, a method is introduced to measure both the scientific and the societal relevance of the Humanities and Social Sciences, by measuring the usage of its main publication form: the monograph. While both the monograph and the field of HSS are under pressure, we saw that there is a considerable interest; from both inside and outside academia. We could say that this is a good result: it indicates the scholarly impact and the societal relevance of HSS. Furthermore, it was possible to measure the influence of subject and language. On the other hand, some of the results were mixed. The usage patterns differ strongly from the literature on societal relevance: contrary to expectations, the data show a low usage percentage by 'professionals'. Whether this is a property of HSS usage or it is caused by the used channel and data set is a question that needs further research. However, the results of this article are promising, and the proposed method can be used as an addition to the existing toolkit.

8 Do developing countries profit from free books? : Discovery and online usage in developed and developing countries compared

Snijder, R. (2013). Do developing countries profit from free books?: Discovery and online usage in developed and developing countries compared. *The Journal of Electronic Publishing*, 16(1), 1–14. https://doi.org/10.3998/3336451.0016.103

8.1 Introduction

The discussion on open access (OA) has many aspects; one of those aspects is the digital divide between developed and developing countries. The digital divide is defined as the inequality in access to the internet, both in a technical sense – a less than optimal infrastructure – and in lack of knowledge to make the best use of the available online resources. Open Access is seen as a way to lower financial barriers for scientists and other readers. Recently, this was discussed by Swan and Hall (Swan & Hall, 2010), who conclude that while putting the idea in practice is not simple, the growth of OA is not only inevitable but also desirable.

Before them, several others also saw chances in freely accessible scientific publications. While each author discusses the inequalities from different angles, the possibilities of open access publishing – combined with changes in institutional and political structures – offer a chance for improvement. Ahmed discusses the digital divide in Africa in great detail, and identifies the required policy changes to amend it. (Ahmed, 2007) Salager-Meyer discusses the inequalities that exist in academic publishing between the developed and the developing countries. Her focus is on journal publishing. (Salager-Meyer, 2008) Christian also discusses the inequalities in funding, IT related infrastructure and possible misconceptions about OA. (Christian, 2008) Likewise, Papin reports on the difficulties that arise when funding is not adequate for publishing in open access journals that charge an author's fee. (Papin-Ramcharan & Dawe, 2006) In the chapter 'Development' Willinsky describes the difficulties that university libraries in developing countries face, and proposes developing and OA publication model as a

possible remedy. (Willinsky & Parry, 2006) Armstrong and Ford focus on the intellectual property rights by discussing the effects of WIPO treaties in contrast to licenses based on Creative Commons. (Armstrong & Ford, 2006) Chan and Costa review several programs such as HINARI, AGORA, eIFL.net, PERI and compare them to directly publishing in open access journals and 'green' open access. (Chan & Costa, 2005) And Ghosh and Kumar Das conclude in their extensive overview that India is leading the open access movement among the developing countries and – by doing so – it is making the developed countries aware of the qualities of scholars and scientists from the developing countries. (Ghosh & Kumar Das, 2007)

Very little research is published on the effects of open access publishing on developing countries, mostly on the citation impact of freely accessible articles. Calver and Bradley investigated citations of OA and non-OA papers in six journals and four books published since 2000, in the field of conservation biology. They did find an OA citation advantage for book chapters, but the number of citations papers or chapters received from authors in developing countries did not increase. (Calver & Bradley, 2010) Norris, Oppenheim, and Rowland, however, did see a larger percentage of citations from developing countries given to OA articles in the field of mathematics than is the case for citations from developed countries. (Norris, Oppenheim, & Rowland, 2008) Walker describes the growth of Bioline International, which enables OA publishing of journals from a wide range of developing countries. Apart from usage data, she describes the citation advantage enjoyed by open access articles. (Walker, 2009)

No research was found on usage of articles, or on academic books.

8.2 Open access monographs and the digital divide

This article tries to answer the question whether open access publishing does actually help to lessen the digital divide between developed and developing countries. As will be described in more detail below, usage data of an earlier experiment with open access monographs was combined with geographical data: from which country does the traffic originate? All countries were divided into two groups: developing countries and developed countries. In order to find whether open access does have a positive effect on developing countries, a group of titles with restricted access was compared to another group of fully accessible monographs. Using statistical analysis, the percentages of book discovery and usage were compared. If the

percentages of the group of fully accessible titles are significantly higher, the claim that open access does benefit developing countries may be supported.

The collection of monographs used for this experiment were published by Amsterdam University Press (AUP), an academic publisher mainly of books in the field of humanities and social sciences. AUP is owned by the University of Amsterdam and works on a not-for-profit basis. (AUP, 2012) AUP publishes around 200 books per year, combined with several journals, some of which are published both on paper and online, some as an open access e-journal. AUP has coordinated the OAPEN (open access Publishing in European Networks) project where several academic publishers worked together to develop an open access business model for monographs in humanities and social sciences, combined with the creation of an open access library. (Open Access Publishing in European Networks, 2010a) From spring 2011, OAPEN continued as a separate organization with AUP as one of its shareholders.

In 2009, an experiment was conducted at Amsterdam University Press to measure the impact of open access publishing of academic books. (Snijder, 2010) During a period of nine months three sets of 100 books were disseminated through an institutional repository, the Google Book Search program or both channels. A fourth set of 100 books was used as control group. As one of the research questions concerned the role of dissemination channels, this division was used.

One of the findings was that open access publishing enhances discovery and online usage of academic books, regardless of the dissemination channel used. Therefore, in this article the titles will be divided into a group of freely accessible titles – without taking into account the dissemination channel – and a closed access group. From April 2009 until December 2009, access to the 400 publications was strictly controlled. Since then, access to several titles has changed which strongly impacts the discovery and online usage.

While the experiment confirmed that books in open access were found more and were used more, it was not known who was using them. The Google Book Search program enabled[5] publishers to monitor geographic information: how many times are books opened from which country? Therefore, in the first months of 2011 this data was gathered and combined with the existing data to answer the research question: *does a change in accessibility of academic books have an effect on developing countries?*

From this question, two hypotheses were derived:

5 This feature became unavailable since the last months of 2011

Hypothesis 1: *The discovery of fully accessible titles in developing countries is significantly higher, compared to titles which are not fully accessible.* Discovery is measured as the number of 'Book visits' a title receives in the Google Book Search program. Book visits are defined as each time that a unique user views a book. (Google Books, n.d.)

Hypothesis 2: *The online usage* (*i.e. pages read*) *of fully accessible titles in developing countries is significantly higher, compared to titles which are not fully accessible.* Online usage is measured as the number of page views a title receives in the Google Book Search program. Page views are defined as the number of unique pages a user views within a 24-hour period. Regardless of the number of times that a unique user views a page, it is only registered once. (Google Books, n.d.)

8.3 Setup of the experiment

The first question to be answered of course is which countries are developing countries. Countries differ wildly in all aspects, and deciding which factors are used to decide which country belongs in what group is not easy. For this experiment, all countries listed under 'Emerging and Developing Economies' in the *World Economic Outlook Database April 2010* are used – with Somalia added to the list. (International Monetary Fund, 2010) The web statistics revealed traffic coming from 179 different countries. Less than a third of those – 48 countries – are marked as developed countries, although those countries generate 70% of the discovery data and 73% of online usage data.

See http://quod.lib.umich.edu/j/jep/images/3336451.0016.103-00000001.txt for the list of all developing countries.

Dividing all countries in two groups is of course a simplification. Doing so enables us to scale down a problem of enormous complexity to a relative simple question. At this point, quantitative data on the effects of open access on the use of monographs is scarce, especially the use in developing countries.

In order to enable further research, the data for the titles are made available in http://quod.lib.umich.edu/j/jep/images/3336451.0016.103-00000002.csv.

The experiment consists of creating 4 equal sets of 100 titles; each title is placed in one of four sets. The different sets are defined using two variables: accessibility and channel: each set is disseminated using a specified channel and accessibility settings. For a period of 9 months, starting in April 2009,

the effect on discovery and online usage is measured. Discovery and online usage are measured using the number of views and downloads from the respective channels.

The division of titles can be summarized as follows:

Table 1 Accessibility of titles

	Set 1	Set 2	Set 3	Set 4
Fully accessible in Google Book Search	No	No	Yes	Yes
Fully accessible through the AUP repository	No	Yes	Yes	No

Set 1: Available 'as usual'. An electronic version of almost all books by AUP is submitted to the Google Book Search website. By default, AUP allows a user of Google Book Search to see only 10% of the book's contents. The full content of each book is indexed by the Google search engine. The titles in this set are not uploaded into the AUP repository. The accessibility of this set is the lowest.

Set 2: Freely available via the repository; visible for 10% in Google Book Search. The titles of this set are uploaded in the AUP repository. For each title, a record is created in the repository database containing metadata and an electronic version of the book. The 'visibility settings' of Google Book Search are not changed and remain at 10 %.

Set 3: Visible for 100% in Google Book Search and freely available via the repository. The titles of this set are uploaded in the AUP repository, and the 'visibility settings' of Google Book Search are set to 100%. The titles in this set are fully accessible through both channels. The accessibility of this set is the highest.

Set 4: Visible for 100% in Google Book Search; not available via the repository. For this set, the 'visibility settings' of Google Book Search are set to 100%. The books are not placed in the AUP repository.

As all titles are accessible through the Google Book Search program, the interest of readers can be measured with the Book Search usage statistics. Here, 'Book visits' – measuring the number of times the webpage of the book is accessed – and 'page views' – which measure the number of pages opened – are the statistics used. Geographical usage data was also available in the Google Book Search program. For each title, a monthly report was downloaded, containing the percentages per country. The statistics per country are measured by applying the percentages to the absolute number of Book visits and page views.

When the statistics from each country per title are known, the percentage of usage coming from developing countries is measured. For instance, the title *The Making and Unmaking of an Industrial Working Class : Sliding down the Labour Hierarchy in Ahmedabad, India* was accessed online from 40 different countries between April and December 2009. This resulted in 418 Book visits and 5115 page views. The number of Book visits from developing countries was 208 and the number of page views from developing countries was 2737. So, for this title the percentage measured for Book visits is 49,8 and the percentage for page views is 53,5. It may not come as a surprise that traffic from India explains the high percentages for this particular title. As I will explain later, this percentage is an exception: for most titles, there is a large gap between usage by developed countries and developing countries.

Furthermore, this paper will not discuss in detail the differences in discovery and online usage from individual countries. For that, several variables that are not part of the data must be examined. One of the criteria is the role of English within the academic communities. While English is widely used, it may not be the preferred language in all communities. The data set does not contain books in French, Spanish, Portuguese or Mandarin, to name a few languages. Another criterion to examine is the subject of the books. While the examined books cover a wide range of subjects, it may be possible that certain research communities would be as interested in the provided titles as others. As described below, subject is one of the elements used in the selection of titles.

8.4 Selection of titles and removal of bias

In April 2009, 893 titles were available at AUP. Using commercial availability, imprint, publication date and series this list was reduced to 412 ISBNs. Of this list, 22 titles were published both as a hardback and a paperback book. As this distinction is irrelevant in the digital domain, 11 ISBNs were removed from the list. Then the oldest title – published in 1994 – was removed, resulting in a list of 400 titles.

Considerable effort has been put in the removal of bias. This experiment operates using four sets of books; therefore, these sets must be as equal as possible. Each of the 400 books is compared using the following criteria: subject; type of work; language; expected sales and publication date.

In the database of AUP each title is assigned several subject codes describing the content. For the sake of the experiment, all titles from a 'subject based series' were assigned the same subject codes. Furthermore,

the number of subject codes was reduced in order to create relatively large groups with the same subject. The same principle was applied to the expected sales, measured by the print run. While each individual title may have a different print run – from 0 for Print on Demand titles to 6500 – an amount rounded up to the next 500 was used. This created again relatively large groups, which could be evenly divided over the four sets. Also, the publication year and the language of the title were taken into account and were 'spread' as evenly as possible.

The 400 titles are written in three languages: Dutch (212 titles); English (180 titles) and German (8 titles). One could argue that using a large percentage of Dutch language titles favours the usage from Belgium and the Netherlands. Furthermore, German is mostly spoken in European countries. For this reason, the selection is reduced to the English language titles only. Language is one of the criteria used to create equal sets, so excluding Dutch and German still leads to a balanced distribution: Set 1 contains 43 titles; Set 2 contains 49 titles; Set 3 contains 42 titles and Set 4 contains 46 titles.

As stated before, previous research found that open access publishing enhances discovery and online usage of academic books, regardless of the dissemination channel used. (Snijder, 2010) Therefore, the statistical analysis will be conducted on the average measurements of the open access channels versus the data from the closed access channel.

8.5 Research results and documenting the digital divide

A first analysis of the data clearly shows the digital divide between developed and developing countries. When looking at the discovery of books, only 30% of the internet traffic comes from developing countries. This is in stark contrast with the United States, from where 19% of all traffic originates. Table 2 depicts the five highest ranking developed and developing countries, and Figure 1 illustrates this.

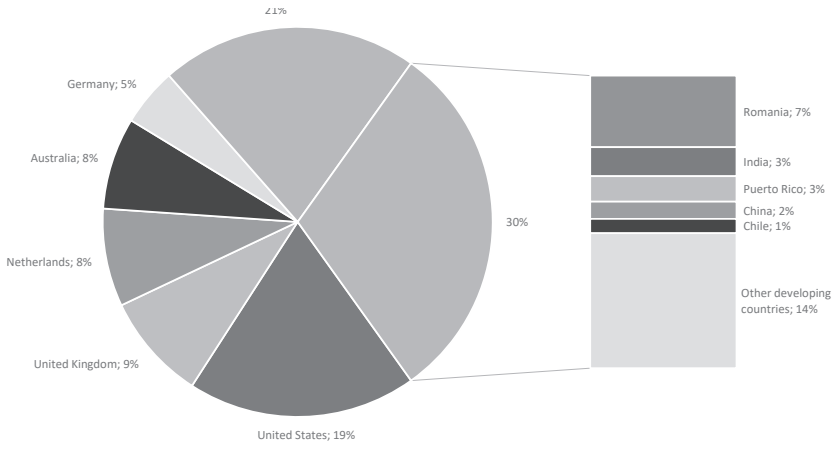
See for a complete list http://quod.lib.umich.edu/j/jep/images/3336451.0016.103-00000003.csv.

Table 2 Discovery and the digital divide: 5 highest ranking developed and developing countries

Country	Discovery: Book visits	Percentage
United States	27262	19%

Country	Discovery: Book visits	Percentage
United Kingdom	12704	9%
Netherlands	11656	8%
Australia	10929	8%
Germany	6904	5%
Other developed countries	30652	21%
Romania	10616	7%
India	4303	3%
Puerto Rico	3781	3%
China	2556	2%
Chile	2126	1%
Other developing countries	20023	14%

Figure 1 Discovery and the digital divide



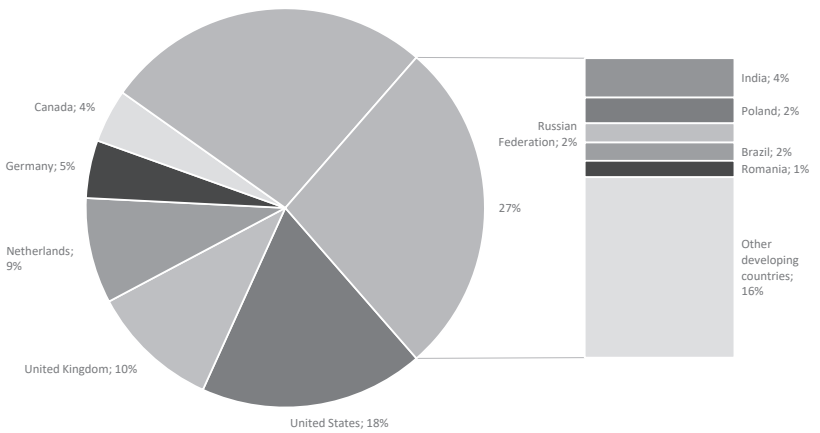
The same story could be told about online usage. Here, the percentages are almost equal: the developing countries account for 27% of the total number of page views, coming from 30% of all Book visits. Again, the country with the largest usage percentage is the United States with 18%. Also, the second and third largest portions come from the same countries: United Kingdom and the Netherlands. See Table 3 for the five highest ranking developed and developing countries and Figure 2 for more details.

See for a complete list http://quod.lib.umich.edu/j/jep/images/3336451.0016.103-00000003.csv.

Table 3 Online usage and the digital divide: 5 highest ranking developed and developing countries

Country	Online usage: page views	Percentage
United States	271130	18%
United Kingdom	155914	10%
Netherlands	127581	9%
Germany	69829	5%
Canada	64606	4%
Other developed countries	396584	27%
India	53235	4%
Poland	34855	2%
Russian Federation	25990	2%
Brazil	24772	2%
Romania	22163	1%
Other developing countries	244135	16%

Figure 2 Online usage and the digital divide



The previous analysis did show the large gap between the developing and the developed countries, but the effect of open access publishing was not taken into account. The experiment run on 400 titles – see (Snijder, 2010)

– revealed that discovery and online usage of books was enhanced. When looking at the 180 English language titles, the same pattern emerges: books that are freely accessible online are found more and used more. This effect can be found with the developed and the developing countries.

Again, the digital divide is also clearly visible. The average number of Book visits is used as measure for the discovery rate of books. When those books are published in closed access, the average rate from developing countries is 127 versus 491 in developed countries; in other words: 20.6% of the average Book visits come from developing countries. Open Access leads to higher average rates: 611 for developing countries versus 1291 for developed countries; the developing countries are responsible for 32.1%.

Online usage is affected in the same way: the average number of page views of books in closed access is 1526 for developing countries and 4542 for developed countries; the percentage for developing countries is 25.1%. Making books fully accessible leads to an average of 5357 page views from developing countries, compared to 14278 page views in developed countries; the percentage for developing countries rises to 27.3%. This is illustrated in Figure 3 and Figure 4.

Figure 3 Discovery and open access

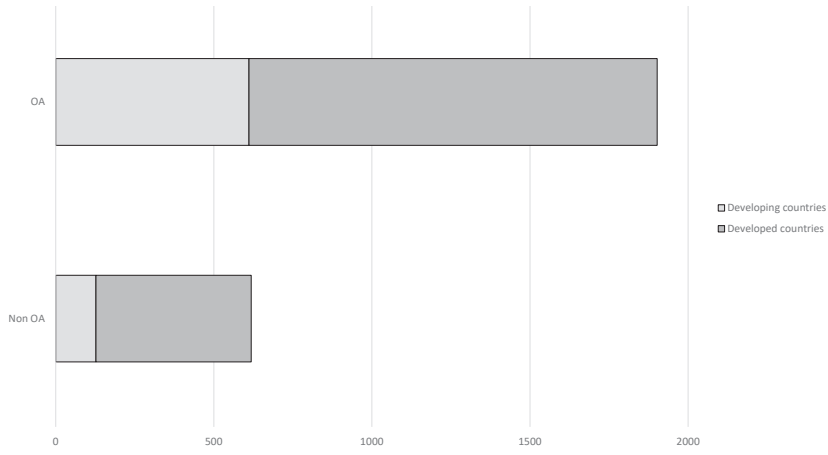
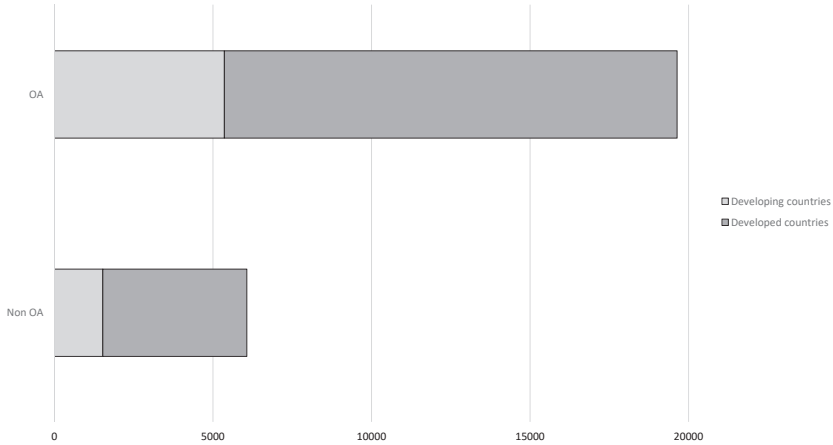


Figure 4 Online usage and open access



The research question 'does a change in accessibility of academic books have an effect on developing countries?' was translated into two hypotheses. The experiment's data was analysed using ANOVA (analysis of variance) in order to test the hypotheses. The results are summarized in Table 4.

Table 4 Hypothesis results

Hypothesis	Result
Hypothesis 1: The discovery of fully accessible titles in developing countries is significantly higher, compared to titles which are not fully accessible.	There was a significant effect of accessibility on discovery in developing countries, $F(3,176) = 1.76$, $p < .05$, one-tailed.
Hypothesis 2: The online usage (i.e. pages read) of fully accessible titles in developing countries is significantly higher, compared to titles which are not fully accessible.	There was a significant effect of accessibility on online usage in developing countries, $F(3,176) = 1.78$, $p < .05$, one-tailed.

8.6 Discussion of the results

Hypothesis 1 states that the discovery of fully accessible titles in developing countries is significantly higher compared to titles which are not fully accessible. Discovery was measured as the percentage of Book visits emerging form developing countries a title received in the Google Book Search program during the experimentation period. The results of the experiment confirmed the hypothesis, which strengthens the claims of

the advocates of open access: access to researchers from the developing countries is improved.

The results are also in line with predictions from the library and information sciences and the field of e-commerce. In the library and information sciences, accessibility to scientific output is linked to research impact. This is discussed by Harnad *et al.* (Harnad *et al.*, 2004, 2008) When barriers are removed, the output – in this case: academic books – is used to maximum effect. The field of e-commerce uses a concept called search costs, which acts as a barrier to transactions. For a more comprehensive discussion of this concept, see Bakos and Granados, Gupta, and Kauffman. (Bakos, 1991; Granados, Gupta, & Kauffman, 2006) Making the complete content of a publication available should lower the search costs considerably, especially if search engines have complete access as well, which may lead to easier discovery of the book. Here, the transaction is acquiring an academic book. Publishing in open access lowers those barriers, which indeed has positive effects.

Book visits are used as an approximation to discovery: it was not possible to measure if a Book visit occurred by a 'new' reader or by a 'returning' reader.[1] Therefore, we cannot state that 204 Book visits from developing countries are equal to 204 new readers of that title. If we assume that a percentage of those Book visits are made by returning readers, the differences in Book visits between titles published in closed access and titles in open access still convey relevant information on the discovery rate. Further research is needed to measure the percentage of new versus returning readers, and whether accessibility influences this.

Hypothesis 2 states that the online usage (i.e. pages read) of fully accessible titles in developing countries is significantly higher, compared to titles which are not fully accessible. The results of the experiment confirmed the hypothesis, which is – again – in line with expectations. Online usage is of course closely linked to the amount of information that is directly available. It should therefore not come as a surprise that making a book fully accessible online leads to more pages read. It is interesting to note however, that the average number of pages read in developed countries is much higher than in developing countries. Presumably the differences in infrastructure play an important role here.

The two confirmed hypotheses refer to data at a high aggregation level; individual countries are not compared for reasons that were discussed

1 According to Google, measured Book visits are done by 'unique users'. It is not clear if this means unique within 24 hours, or any other period of time.

earlier: the role of available languages in the data set and the diversity of subjects. While bias has been removed as much as possible, the data set may be relatively small. One could consider the contents of the OAPEN Library, containing hundreds of titles published by dozens of different publishers. However, this collection does lack a control group, making it harder to draw conclusions based on its performance. On the other hand, research on freely accessible books by Hilton and Wiley was done on 41 titles, of which 7 were non-fiction books. (Hilton, 2011)

In order to enable further research, the data for all 180 titles are made available in http://quod.lib.umich.edu/j/jep/images/3336451.0016.103-00000002.csv.

8.7 Conclusions

In the introduction, the technical and cultural barriers to the use of open access were discussed. Online access to information resources does require an infrastructure that supports it. Furthermore, lack of knowledge or cultural biases may impede the usage of OA. Because of the way this experiment was set up, these factors do not play a significant role. The technical requirements for finding and using the titles from both the freely accessible group and the control group were exactly the same: all were available through the same dissemination channel: the Google Books Search program. Also, all titles were available in the same time period. The non-technical barriers may have played a role, but if that was the case their influence would be the same on all books. As we have seen in the discussion of the selection of the titles, much effort is placed in the removal of bias. Therefore, the group of openly accessible books is balanced with the control group.

Research on the effects of free online accessibility of books is scarce, especially the effects on academic books. As open access is gaining momentum as dissemination model – see for instance the briefing paper of Knowledge Exchange – there is greater need for knowledge of the effects it has on all stakeholders, both in developing and in developed countries. (Knowledge Exchange, 2010) The findings of this article reaffirm the notion that removing barriers to access has positive effects on discovery and online usages of academic books. This is beneficial for researchers from both developing and developed countries, and it does indicate that open access makes it possible to "[…] share the learning of the rich with the poor and the poor with the rich, [making] this literature as useful as it can be […]". (Chan *et al.*, 2002)

Furthermore, the data used reflects the situation of 2009. As described by UNESCO, several developing countries are investing heavily in Research and Development. (UNESCO, 2010) This will impact the discovery and online usage of academic books, and it will be interesting to see if the digital divide becomes smaller in the next few years.

9 Revisiting an open access monograph experiment : Measuring citations and tweets five years later

Snijder, R. (2016). Revisiting an open access monograph experiment: measuring citations and tweets 5 years later. *Scientometrics*, (May), 1–19. https://doi.org/10.1007/s11192-016-2160-6

9.1 Introduction

While the question whether publishing in open access (OA) leads to a citation advantage has been studied numerous times for journal articles, much less work has been done in the realm of monographs. This imbalance is further illustrated by the fact that literature on articles is listed in several overviews – for instance by Archambault *et al.* (2014; 2016) or SPARC Europe (2015) – while publications on monographs are scarce.

The impact of scholarly publications has traditionally been assessed through citations, and, more recently, altmetrics have come into use as another type of impact measure. Here, altmetrics are defined as the measurement of online activities about scholarly publications. A specific form of altmetrics – Twitter mentions – will be used as an indicator of societal rather than academic impact of scholarly books.

Until recently, books have been largely ignored by those attempting to measure impact: both in the realm of citations and altmetrics. This paper will address this lacuna by analysing the role of open access on the impact of books, based on experimental data.

In 2009, an experiment was conducted on 400 monographs, measuring the effects of open access on discovery, online consultation, sales figures, dissemination channels and citations (Snijder, 2010). In line with expectations, the experiment found that making books freely available enhances discoverability and online consultation. Furthermore, no significant influence on sales could be established. These outcomes are consistent with the results of other investigations (Ferwerda *et al.*, 2013; Snijder, 2014b).

The experiment was less successful in establishing whether OA enhances the scholarly impact of books in a more traditional sense: through citations.

Revisiting the experiment will help to answer this question. At the conclusion of the 2009 experiment no citation advantage for freely accessible books could be found. This is in contrast to journal articles, where higher citation rates for OA have been frequently reported. In October 2014 citations of the 400 monographs included in the original experiment were measured again, this time combined with the number of tweets mentioning each book.

In 2009, it was not possible to assess whether making monographs freely available enhanced scholarly impact, nor could anything be said about influence on society at large. This paper revisits the experiment, drawing on additional citation data as well as developments in the altmetrics landscape. It attempts to answer the following research question: does open access have a positive influence on the number of citations and tweets a monograph receives, taking into account the influence of scholarly field and language? Furthermore, looking into the correlation between monograph citations and tweets helps to determine whether these measurements are related.

9.2 Background

This review focuses on monographs, starting with monograph citations before discussing alternative impact metrics as they relate to books. Apart from availability, two other factors may influence citations and altmetrics uptake: scholarly field and language. Different citation cultures exist within individual fields of study, making it hard to compare bibliometrics data between disciplines without normalisation.[1] Furthermore, some authors suspect a bias towards English language publications in citation databases; this will be discussed in section *The influence of language*. The language of the publications included in the experiment discussed in this paper may affect its outcomes, as roughly half of the books included in the study – 178 books – are written in English; the remaining 212 books were written in Dutch or other languages.

Another recurring theme in the literature on OA is the correlation between citations and altmetrics, see for instance Thelwall *et al.* (2013). If altmetrics are closely connected to scholarly impact, one might expect a statistically significant correlation between them. On the other hand, when altmetrics are seen as measuring a different type of interest in scholarly output – rather than as a proxy for citations – it may be more useful to search

1 See for instance section *1.2.3. Normalisation of citation impact indicators* of Wouters *et al.* (2015)

for online activity relating to scholarly books with the weakest correlation to citations. In that way, the broadest possible spectrum of engagement with monographs may be captured. This is discussed further in section *What is the relation between citations and altmetrics?*

One of the assumptions of the original 2009 experiment was that making monographs available in open access enables more researchers to read books that would otherwise be inaccessible. The results of the experiment pointed to significantly greater usage – discoverability and online consultation – for freely available books. It was assumed that enhanced access would also lead to more citations, as it made books available to scholars working in more restrained environments. This assumption is challenged by the findings of a survey of 2,231 humanities and social science researchers based in the United Kingdom: only ten percent of the respondents reported having difficulties in accessing monographs (OAPEN-UK, 2014).

This perception by researchers opens interesting possibilities. If professional users of monographs have no serious problems in accessing them, we would expect to find a smaller citation advantage for OA books, or none at all. However, among the outcomes of the 2010 experiment was the improved discovery and online consultation of free online books. We might assume that a significant part of that online usage is coming from readers other than academics. In the discussion of altmetrics outlets, tweets are strongly associated with the wider public (Bornmann, 2014; Haustein *et al.*, 2014). For readers not connected to universities with large library collections, open access has direct benefits, potentially leading to more mentions on Twitter and the wider dissemination of research.

The OAPEN-UK project also looked into researchers' attitudes towards making their books freely accessible. It concluded that authors see open access publishing as a way to increase their readership, and that this perceived benefit of open access is valued by many. However, opinions differ about the way it should be implemented (Collins & Milloy, 2016).

9.2.1 Citations and books

Glänzel & Schoepflin (1999) discussed the differences in citation behaviour in the humanities and social sciences compared to the sciences. They matched the percentage of cited articles to citations to books and other long form publication. In scientific fields such as immunology or solid-state physics, the number of cited articles is over 85%. In contrast, scholars in the fields of sociology and history and philosophy of science tend to cite a much

lower percentage of articles: 40% or lower. In other words: book citations are strongly linked with the humanities and social sciences.

Several researchers have investigated book citations. Tang (2008) analyses citations of 750 randomly selected monographs in the humanities and the sciences. Within each discipline, he finds differences in the number of uncited books, the time span in which half of the citations are occurring, and the recency of citations. In general, the fields of science tend to have lower numbers of uncited books and more recent citations compared to books in the humanities. However, the citation culture within each scholarly field is quite different. Nederhof (2011) deems the results of the impact investigations more useful, when a "citation window" of at least six to eight years is used. According to Nederhof, this better reflects the world-wide reception of the publications. Another factor – not explicitly mentioned by Nederhof – is the fact that writing a book takes considerable more time than writing an article. This might have consequences for the citations in scholarly fields where monographs are the dominant publication form. Using a longer period to accumulate citations in the field of humanities is a solution also proposed by Linmans (2009). By doing so, Linmans is able to assess humanities publications. Furthermore, he expects Google Scholar to be a very useful source of book citations.

The use of Google Scholar as source of citation data is described by Harzing and Van der Wal. By comparing the coverage in the area of management and international business by Google Scholar and Thomson ISI Web of Knowledge, they conclude that Google Scholar is more comprehensive – especially in the area of books and non US journals (Harzing & van der Wal, 2008). Whether Google Scholar or Google Books fares better than Scopus citations is tested by Kousha *et al.* (2011). Based on a set of 1,000 books, these authors determine that the larger amount of citations by the Google products could be used for assessing the publications in book-oriented disciplines in the British humanities and social sciences. More recently, Prins *et al.* investigated the coverage of social sciences and humanities by Web of Science (WoS) and Google Scholar. They conclude that the coverage by Google Scholar is better for these scholarly fields, although the quality of the data is not as consistent as WoS (Prins, Costas, Leeuwen, & Wouters, 2014). In this paper, citations are derived from Google Scholar.

The availability of citation data for monographs is currently not on the same level as articles: the Thomson Reuters' Book Citation Index was first published in 2011, providing citation information relating to a selection of just 25,000 titles (Jump, 2011). The paucity of citation data relating to books within the prominent citation databases has inspired several authors to

explore alternative sources of citation information. For instance, Kousha and Thelwall use the Google Books index to identify citations from books. Their goal is to compare the number of citations in the Thomson Reuters/ Institute for Scientific Information databases (ISI) to those in Google Books. It is interesting to note that the ratios strongly differ between scholarly fields (Kousha & Thelwall, 2009). This is in line with the conclusions of Nederhof, discussed earlier in this paper. Recently, Thelwall and Sud have used the Thomson Reuters Book Citation Index (BKCI) to explore whether co-authorship of monographs leads to a higher citation impact. Contrary to the results found for articles, the authors conclude that co-operation does not generally lead to more citations (Thelwall & Sud, 2014). Again, we see that citation behaviour for monographs differs from journal articles.

9.2.2 Altmetrics

In the document "altmetrics – a manifesto", altmetrics are described as an additional dimension to complement citation data. As publications are made available on the web, usage can be measured immediately (Priem *et al.*, 2011). The online activities considered within altmetrics frameworks are diverse: a non-comprehensive list includes blog posts; tweets; Scopus citations; CiteULike bookmarks; Mendeley references or Facebook posts. The question of whether altmetrics measure usage from the academic world or should be treated of an indication of interest from wider reading communities will be discussed in the next section: *What is the relation between citations and altmetrics?*

In the realm of monographs and other book-length publications, several researchers have been working on alternative ways to assess scholarly value. Perhaps not surprisingly, data from academic libraries is used. For instance, White *et al.* discuss 'libcitations', where the number of academic libraries holding a certain book is the unit of measure. The collection of a library is based on qualitative decisions; a monograph that is acquired by a large number of libraries has a larger impact compared to a monograph that only resides in a few libraries. The authors do not compare those metrics to citation data (White *et al.*, 2009; Zuccala & White, 2015). In contrast, Cabezas-Clavijo *et al.* (2013) use the number of library loans from two academic libraries as a proxy of scholarly impact. When the library-generated data is compared with the available citation data, again the same pattern emerges: at best a weak correlation between the 'alternative' metrics and citations. Quite a different approach is used by Zuccala *et al.* (2014), who use machine-learning techniques to automatically classify the conclusions

of book reviews in the field of history. However, the reported results derive from a pilot experiment, and no correlation to citations is described.

The question remains which altmetrics outlet to use to assess monographs. Here we face an additional complication: most altmetrics tools use an online unique identifier attached to a publication. In the case of journal articles, this will most likely be the Digital Object Identifier (DOI). Books are usually identified by an ISBN, but the use of ISBNs as digital identifier is not as widely spread as DOIs. This is especially true for the books in our data set. Another aspect to consider is the preferred outlet: are mentions of books evenly spread among all outlets? If that is not the case, which outlet or outlets are to be measured? Hammarfelt (2014) has compared the coverage in several online sources of 310 English language articles and 54 books – also written in English – in the field of humanities and social sciences. He concludes that for books, Twitter delivers the most results. In order to identify books, the title – or a significant part of the title – has been used.

9.2.3 What is the relation between citations and altmetrics?

The relation between citations and altmetrics is currently under investigation. If these measurements are strongly correlated, they might measure something similar. However, if there is no strong connection, can altmetrics be considered to be an indication of a new aspect of impact? The literature discussed in this section is focused on journal articles; the connection between citation, altmetrics and books is poorly researched and there is little existing literature on the topic.

Several large-scale studies on correlations between citations and altmetrics have been performed. Using a set of over 24,000 open access articles published in the Public Library of Science, Priem *et al.* (2012) find a large uptake in at least one source of altmetrics activity. Yet, the correlation between citations and altmetrics is not very strong. Costas *et al.* (2014) arrive at a different conclusion regarding altmetrics activity: between 15 and 20% of the articles in their set – based on more than 718,000 publications covered in the Web of Science – are mentioned via an altmetrics outlet, compared to almost 80% in the case of Priem *et al.* (2012), who examined open access articles. Again, they do not find a strong connection between altmetrics and citations.

After a meta-analysis of seven studies, Bornmann (2014) concludes that different types of online outlets vary in the amount of correlation with citation counts. The bookmark counts of online reference managers

Mendeley and CiteULike are the most connected to citations. In contrast, Twitter citations seem to measure something different from traditional citations: the correlation with traditional citations for the number of tweets is negligible. This is also described by Haustein *et al.* (2014), who conclude that Mendeley is predominantly used by the academic community, while Twitter is used by a general audience.

The report by Wouters *et al.* (2015) provides an overview of the current literature on the role of citations and altmetrics in research assessment. The report describes citations and altmetrics as complementary measures which should be considered within the context of the publication. In a recent article by Thelwall (2016) the correlation between citations and altmetrics is also something to be considered within a certain context. Interpreting the correlation strength is quite complicated, as factors such as the average and the variability of the number of citations the documents received tend to play an important – but not always straightforward – role.

9.2.4 Twitter as research tool

Using the number of tweets as an indicator of impact has several advantages when we look at the research at hand. Twitter is a widely-used platform, which has been available since 2006. Due to its global usage and the extended period that it has been available for, we might expect more 'success' in identifying tweets about the books in our data set. The books in the data set analysed during this experiment were published between 1995 and 2008. The relatively long period between the publication of the books studied and the analysis carried out for this paper conforms to the longer 'citation window' discussed by Nederhof and Linmans. It may also allow for the accumulation of more tweets, which seems to be the case here. Moreover, Hammarfelt (2014) describes Twitter as the platform containing the most mentions of books, compared to other sources of altmetrics data. In the paper by Hammarfelt, the highest number of tweets for one book was 19. In our data set, 48 of the 400 books were mentioned in 25 tweets or more.

The results for this paper were derived using a search tool. While Twitter.com has its own search engine, a sample test performed in October 2014[2] indicated that Topsy.com was more successful in identifying tweets about the books in the data set. This search engine had indexed all publicly available tweets, making it a serious alternative to the Twitter.com search

2 The Topsy.com service has been discontinued in December 2015.

engine (Sterling, 2013). Therefore, Topsy.com was used to identify relevant tweets for the purposes of this study.

9.2.5 The influence of language

Little research is available on the influence of language on monograph citations. Abrizah and Thelwall (2014) have investigated – among other influences – the role of language in the number of citations Malaysian monographs received. While 71% of the books analysed were published in Malay and the rest in English, the English language books were significantly more likely to be cited. Again, the authors have found differences between the citations in the different scholarly fields. Other researchers investigated the role of language on the citation rate of articles, by comparing the 'native' language to English (Aleixandre-Benavent *et al.*, 2007; Guerrero-Bote & Moya-Anegón, 2012; Winkmann, Schlutius, & Schweim, 2002). The common factor here is the bias of citation databases towards English, which disadvantages articles in other languages.

The relationship between language and Twitter usage has also been investigated. The paper by Hong *et al.* (2011) reveals the large proportion of English language tweets in the examined data set of over 62 million tweets. The number of tweets in English consist of 51% of the total. This may also affect our outcomes, and we might expect more tweets for books in English, compared to the books in Dutch.

9.2.6 The influence of subject

Nederhof (2011) describes citation impact measurements in modern language and linguistics research. Although these fields are closely connected, there are significant differences in publication and citation behaviour within each field. Whether the differences in citation patterns is also reflected in the number of tweets relating to books in different subject fields is not clear.

Holmberg and Thelwall (2014) examined a related question, by looking at disciplinary differences in how researchers use Twitter. This research was centred on all the tweets by scientists in ten disciplines. In contrast, this paper only examines tweets that mention the books in our data set. Holmberg and Thelwall conclude that differences in Twitter usage exist between scientific fields: those working on biochemistry, astrophysics, cheminformatics and digital humanities use it for scholarly communication. Others, who specialise in economics, sociology and history of science, are not deploying the microblogging site for their work. No information about

the affiliation of the Twitter users in our data set is available, which makes it difficult to replicate this type of research.

9.3 Research setup and the data set

The Introduction discussed whether publishing in open access has a significant effect on the scholarly impact of monographs, using citations and tweets. However, based on the literature review, we might expect additional influences by the scholarly field and language. Language is an important factor, as half of the collection analysed in this experiment is in Dutch, while the other half consists mainly of English language books. The study attempts to answer the research question, while taking into account these influences. Furthermore, we might expect a loose correlation between the number of citations and altmetrics. This is another aspect to be examined in this paper.

The data set consists of 400 books, all published by Amsterdam University Press (AUP), in the period 1995 to 2008. In the original experiment the books were divided into 4 sets of 100 titles (Snijder, 2010). Three sets were immediately made available in open access; the fourth set was used as control and lacked full online availability. The books in the experimental data set were made available without embargo. Since the end of the experiment, the publisher has changed the availability of several books. The changes in availability since 2009 explain the percentages of OA in our data set: instead of 75%, 68% of the books are now freely available.

In the data set, 22 different subjects can be identified; in this data set we will treat the subject of the books as a proxy for scholarly field. The subjects are not evenly spread over the books: while 25% of the titles discuss public administration and political science, the combination of the six topics education, economics, mathematics, theatre, information technology and religion accounts for just 6% of the books.

In order to create groups of comparable size, the books were placed in two subject-based groups. Books on the subjects Archaeology, Art - History, Culture, Dutch Language, Education, History, Japan, Law, Literature, Motion Pictures, Music, Philosophy, Religion, and Theatre were included in the "Humanities" group. Books on Economics, Information Technology, Mathematics, Medicine, Psychology, Public Administration and Political Science, Science, and Sociology were placed in the "Other scholarly field" group.

Table 1 Books in data set broken down for availability and subject

	N	Percentage		N	Percentage
Accessibility			**Scholarly Field**		
Open access	271	68%	Humanities	138	35%
			Other scholarly field	133	33%
Non OA	129	32%	Humanities	82	21%
			Other scholarly field	47	12%

Compared to the number of subjects, the number of languages is quite small. More than half of the data set – 212 books – comprises books published in Dutch; 178 books are published in English, while the remaining group of ten books are in either German or dual-language English-Dutch books. For the purposes of this study's analysis, the books are divided in English-language titles and titles in other languages. The background section discussed the role of English; given the fact that only ten of the remaining books were not written in Dutch, they were not placed in a separate group.

Table 2 Books in data set broken down for availability and language

	N	Percentage		N	Percentage
Accessibility			**Language**		
Open access	271	68%	English	129	32%
			Other languages	142	36%
Non OA	129	32%	English	49	12%
			Other languages	80	20%

Table 3 lists the combined data: the books divided into 8 groups.

Table 3 Books in data set broken down for availability, subject and language

	N	Percentage		N	Percentage
Accessibility			**Subject and language**		
Open access	271	68%	Humanities – English	66	17%
			Humanities – Other languages	72	18%
			Other scholarly field – English	63	16%
			Other scholarly field – Other languages	70	18%
Non OA	129	32%	Humanities – English	22	6%
			Humanities – Other languages	60	15%
			Other scholarly field – English	27	7%
			Other scholarly field – Other languages	20	5%

For complete details, please see the data set, available at http://dx.doi.org/10.17026/dans-x6m-67b2.

As mentioned in section *Citations and books*, the source of citations chosen for the purposes of this study is the Google Scholar website. In 2009, the citations were measured during the month August; in 2014, the citations were assessed in October. In the results section of this paper, the differences in citations will be discussed in more detail.

Most altmetrics tools use online identifiers – such as DOIs – to identify journal articles. Identifying publications turns out to be more problematic for monographs, which are more commonly associated with an ISBN. In contrast to DOIs, ISBNs are not widely used as an online identifier and searching for tweets using ISBNs did not prove to be successful. In contrast, searching for tweets using book titles delivered more results. Furthermore, personal communication with the founder of Altmetric.com has confirmed that – at the moment of writing – no online identifier for monographs can be used.[3] In other words, a stable online identification was not available. However, searching for tweets using titles has disadvantages, particularly in relation to books that have been published in several editions. As far as could be established, the books in the data sets have not been published in several editions.

3 Euan Adie, personal communication, 10 February 2014

Apart from availability in open access, language and scholarly field have also been identified as having a possible impact on the number of tweets relating to any given title. According to the paper by Hong *et al.* (2011) half the tweets of their set containing 62 million tweets are sent in English. This may affect the number of tweets about books written in English as well. About half the books in our data set were written in Dutch. We can assume that these books will be read more by people in Dutch-speaking countries, while books written in English may attract a more global audience. Secondly, we have seen that within each scholarly field, the citation patterns are different. Does a comparable divide also exist in the use of social media? Are some subjects more prone to attract tweets than others? In our investigation, we will take both language and subject into consideration, combined with open access.

9.3.1 Obtaining citations using Google Scholar

The citations were obtained using the same method as in the original experiment (Snijder, 2010). For each of the monographs a URL pointing to a search at Google Scholar was constructed. The URL was based on the main title, placed between quotes. If needed, parts of the subtitle or the name of the author were added to ensure a best possible match. For instance, to find the book *Why Are Artists Poor? : The Exceptional Economy of the Arts*, by Hans Abbing, the following URL was used: http://scholar.google.com/scholar?hl=en&lr=&q=%22Why+Are+Artists+Poor?%22+The+Exceptional+Economy. This automatically opens the English language interface of the Google Scholar website, using the following search query: *"Why Are Artists Poor?" The Exceptional Economy*. Using quotes forces the website to search on the exact phrase; in this case, part of the subtitle was added to narrow down the results. The resulting number of citations was recorded.

The search was done manually, over several days. Restrictions on the Google Scholar website limit the number of searches that can be carried out within a short time. As was the case with searching for tweets, using the book title instead of the ISBN yielded the best results. The data set for this paper contains the search queries used on both Google Scholar and the Topsy.com website.

Each result was examined critically, and when multiple instances of a title – each containing their 'own' number of citations – were found, only the result with the highest number of citations was used. In the example below, the number of recorded citations was 25, not 29 (25+2+2). This method was also used in the experiment carried out in 2009.

- [BOOK] Reformation of Islamic thought: a critical historical analysis NḤA Zayd - 2006 - books.google.com Cited by 25
- [CITATION] Reformation of Islamic Thought. A ritical Historical Analysis, wrr-Verkenning nr. 10 N Abu Zayd - 2006 - Amsterdam: Amsterdam University … Cited by 2
- [CITATION] Nasr,(2006), Reformation of Islamic hought: A Critical Historical Analysis, WRR/Den Haag A Zeyd - Amsterdam University Press, … Cited by 2

9.3.2 Finding tweets using Topsy.com

The method used to find tweets resembles the routine for obtaining citations: again, the book titles were used in a manual process. In order to narrow down the results, quotes were used. For instance, the search for "The Rise of the Cult of Rembrandt" resulted in this URL: http://topsy.com/s?q=%22The%20Rise%20of%20the%20Cult%20of%20Rembrandt%22. Each query was set up in this way. If the search was not successful, the quotes were removed in an attempt to widen the search. All the resulting tweets were examined, and tweets on other subjects than the book in question were not counted. As mentioned before, neither ISBN nor another online identifier have been used. If readers tweeted a link to a book without using the title, this was not recorded.

9.4 The results

As discussed earlier, this paper engages with the following research question: Does open access have a positive influence on the number of citations and tweets a monograph receives, taking into account the influence of scholarly field and language? Additionally, the correlation between the number of citations and altmetrics activity has previously been investigated in relation to journal articles. Here, the correlation between monograph citations and tweets is investigated. Are they connected, or do these figures describe different aspects of impact?

To get a sense of the way that citations and tweets are distributed across our data set, frequency has been plotted in two charts below. In both charts, it is evident that distributions are skewed: most books received between one and five citations or tweets.

Figure 1 Frequency of citations, measured October 2014

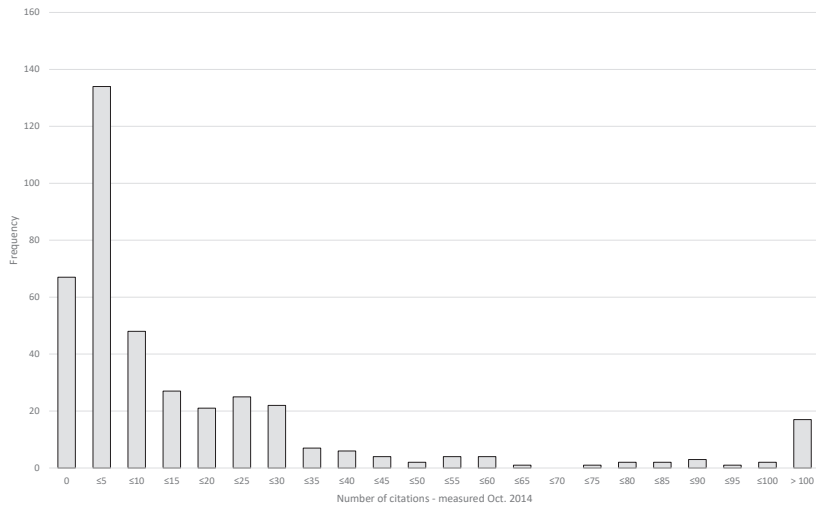
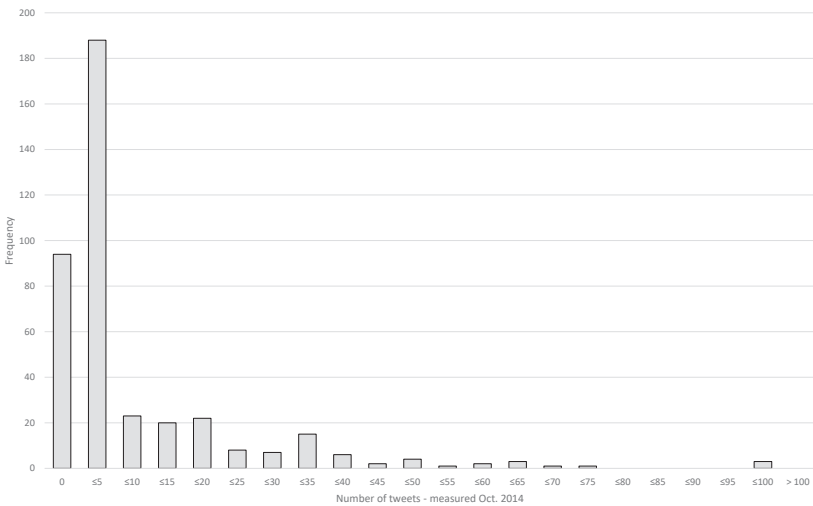


Figure 2 Frequency of tweets, measured October 2014



The books in our data set were published between 1995 and 2008. The literature cited in the background section suggested that a 'citation window' of six to eight years is preferable when assessing monographs. The effect of a longer period on the number of citations is clear in our data set. In 2009 – the period of the original experiment – the average number of citations was 9.0.

In 2014, the average has ascended to 39.0; more than four times as many. Perhaps more telling is the fact that in 2009, 183 of the 400 books received no citations. In comparison, the number of titles without citations has shrunk in 2014 to 67 books, meaning that 83% of the books in the data set had been cited at least once.

The paper by Thelwall *et al.* (2013) predicts the opposite effect for altmetrics: due to a rapidly increasing uptake, newer publications will be mentioned more than older publications. However, Twitter was founded in 2006 – more than a decade after some of the books in the sample examined for this project were published. Twitter analysis for this project was carried out in 2014. Although Topsy did capture historical tweets, the gap between the publication of many of the books in the sample and the advent of Twitter may partly explain the relatively low percentage of books with at least one mention on Twitter: 77%.

Table 4 Books in data set broken down for subject: citations and tweets

Subject	N	Median (Standard deviation)			Books with citations in 2014 (percentage)	Books with tweets (percentage)
		Citations 2009	**Citations 2014**	**Tweets**		
Humanities	220	0 (16.9)	4 (39.8)	2.5 (16.9)	157 (80%)	172 (78%)
Other scholarly fields	180	1 (103.8)	7.5 (211.7)	2 (12.8)	176 (98%)	134 (74%)
Total	400	1 (70.9)	5 (145.3)	2 (15.2)	333 (83%)	306 (77%)

The literature discussed earlier in this paper not only predicts more citations from a longer citation window, but it also describes differences in citation culture. Different citation patterns for different disciplinary areas are clearly visible in *Table 4*. The differences are also visible when books are categorised according to publication language in *Table 5*. The total number of citations counted in 2014 includes citations identified during the 2009 study; the column "tweets" lists the total number of tweets in which each monograph is mentioned.

Table 5 Books in data set broken down for subject: citations and tweets

Subject	N	Median (Standard deviation)			Books with citations in 2014 (percentage)	Books with tweets (percentage)
		Citations 2009	**Citations 2014**	**tweets**		
English	178	2 (104.8)	13 (213.5)	5 (15.6)	158 (89%)	153 (86%)
Other languages	222	0 (14.2)	3 (31.8)	1 (13.9)	175 (79%)	153 (69%)
Total	400	1 (70.9)	5 (145.3)	2 (15.2)	333 (83%)	306 (77%)

The literature on language and citations states that publications in English tend to receive more citations than texts in other languages. This may be in part because citation databases are more likely to index English language databases. In this paper, the citation data is not derived from a database such as Thomson ISI Web of Knowledge, but from Google Scholar. The sources indexed by Google Scholar are not known, and as such it is impossible to assess the extent to which Google Scholar citations are biased towards English language publications.

The influence of language – and the dominance of English – in relation to Twitter usage has also been discussed. In our data set, English language books are mentioned 13.2 times on average, while the average for books in other languages is far lower. Based on this, it seems reasonable to assume that the higher mean for English language books could partly be explained by the number of tweets in English.[4]

9.4.1 Analysis of citations and tweets

To assess the relation between open access, language and subject, the data gathered has been analysed using a generalised linear model (GLM) analysis – in this case a negative binomial regression analysis. This type of statistical investigation allows for response variables that have error distribution models other than a normal distribution. We have seen that the distributions of both citations and tweets do not follow a neat 'bell curve', but are severely skewed. The GLM analysis is used to quantify the strength of the effect of the factors on the number of citations. Here, the factors are accessibility, language and scholarly field.

4 The dataset used does not record the language of the tweets.

9.4.1.1 Citations

The average number of citations for books published in OA was 35.7 (*SD* = 174.4); the mean number of citations for books not made available in OA was 13.4 (*SD* = 36.6). For the total set, the mean number of citations was 30.9 (*SD* = 157.44). Based on this, we might conclude that making books freely available has a large positive effect on the number of citations. If no further statistical analysis is deployed, the conclusion could be that the experiment has produced the expected result. This is also supported by the results of a negative binomial (maximum likelihood estimate) regression analysis. The estimated effect size Exp (B) with 95% Confidence Interval (CI) is listed in *Table 6*. If only accessibility is taken into consideration, making books available in open access leads to 2.6 more citations (8%)[5] on average, compared to those published in closed access.

Table 6 Negative binomial regression: citations

	Exp (B)	95% CI	
Accessibility (Reference = Non open access)			
Open access	2.588*	1.802	3.717
Intercept	14.884*	11.043	20.061

* Significant on 95% level

However, when the effects of language and scholarly field are analysed, the results are more nuanced. *Table 7* lists the results. When controlled for language and scholarly field, making a book freely available leads to 1.7 (5%) more citations on average. However, the 'citation advantage' for books in English is 3.5 (11%) and books in the humanities receive 0.5 citations on average (2%), compared to books on other scholarly fields. The results still point to a slightly positive influence of open access on the number of citations, but the effects of language and scholarly field are also significant.

5 Throughout section 4, the average number of citations/tweets of that data set is used as reference.

Table 7 Negative binomial regression: citations, language, scholarly field

	Exp (B)	95% CI	
Accessibility (Reference = Non open access)			
Open access	1.657*	1.168	2.352
Language (Reference = Other languages)			
English	3.509*	2.529	4.869
Scholarly field (Reference = Other scholarly fields)			
Humanities	0.538*	0.391	0.740
Intercept	12.757*	8.920	18.243

* Significant on 95% level

9.4.1.2 Tweets

The average number of tweets for books published in OA was 9.1 (*SD* = 15.4); the mean number of tweets for books not made available in OA was 7.6 (*SD* = 14.6). For the total set, the average number of tweets was 7.86 (*SD* = 16.044). Again, at a first glance we see an advantage for OA books and we might be tempted to conclude that publishing monographs in open access leads to a higher uptake by social media, in this case Twitter. Nevertheless, this conclusion is refuted by the results of a negative binomial (maximum likelihood estimate) regression: when only open access is considered, the results are not statistically significant.

Table 8 Negative binomial regression: tweets

	Exp (B)	95% CI	
Accessibility (Reference = Non open access)			
Open access	1.188	0.806	1.751
Intercept	6.977*	5.068	9.605

* Significant on 95% level

The results of *Table 9* show that the effects of language and scholarly field are statistically significant, in contrast to accessibility. Books in English receive 2.5 (31%) more tweets and books in the humanities get 1.8 more tweets (22%) on average.

Table 9 Negative binomial regression: tweets, language, scholarly field

	Exp (B)	95% CI	
Accessibility (Reference = Non open access)			
Open access	1.211	0.827	1.772
Language (Reference = Other languages)			
English	2.454*	1.697	3.549
Scholarly field (Reference = Other scholarly fields)			
Humanities	1.779*	1.224	2.585
Intercept	3.032*	1.929	4.766

* Significant on 95% level

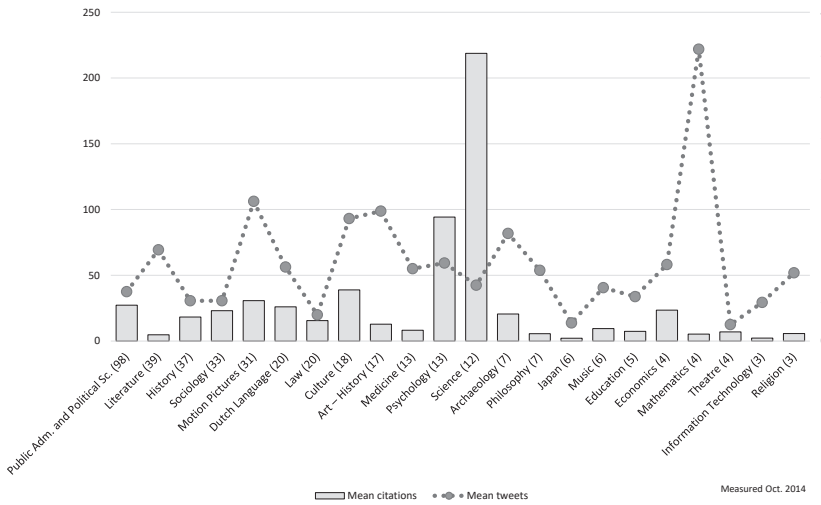
Based on the literature, we might expect that both subject and language are significant factors, whether or not the books have been made available in open access; if different scholarly fields have different citation cultures, this should affect the outcomes. The results point to the same effect on tweets. Yet, we could argue that analysing citations from different scholarly fields is comparing apples and oranges: within each discipline, the average number of citations is different. This may have impacted the results of the analysis. In order to compensate for discipline variance, it is necessary to compare the number of citations and tweets within a group of books with the same subject. The results of this analysis are described in the next section.

9.4.2 Statistical analysis within subject

The books in the data set are not evenly distributed across subjects. While 98 books discuss the subject "Public Administration and Political Science", there are also groups of just three or four books on subjects such as "Economics", "Mathematics", "Theatre" or "Religion".

If the mean number of citations and tweets are plotted on a chart, the differences become visible in a literal sense: the mean number of citations differs between scholarly fields, and a high number of mean citations is not matched by a high number of tweets. The chart also lists the number of books per subject. Whether the results of analysing subject-based groups containing as little as four or three books have any statistical significance is highly doubtful.

Figure 3 Mean citations and tweets - per subject



For this reason, only the five largest subject-based groups are analysed, again using the negative binomial procedure. The following subjects were examined using this approach: “Public Administration and Political Science”, “Literature”, “History”, “Sociology” and “Motion Pictures”. The total number of books in the five largest subject-based groups is quite large: 238 titles. Of those titles, 172 were published in open access, and 66 were not made openly available.

Table 10 The five largest subject-based groups: number of titles, citations and tweets. Measured October 2014

Subject	Open access books			Non open access books		
	N	Median citations (SD)	Median tweets (SD)	N	Median citations (SD)	Median tweets (SD)
Public Administration and Political Science	82	10.5 (40.6)	2 (8.6)	16	6 (95.5)	2 (24.6)
Literature	19	2 (7.0)	1 (28.4)	20	2 (13.0)	0.5 (27.3)
History	22	4.5 (58.4)	2.5 (17.8)	15	4 (10.0)	1 (5.8)
Sociology	22	18 (31.7)	0.5 (8.5)	11	7 (18.5)	2 (11.0)
Motion Pictures	27	24 (44.5)	15 (13.8)	4	21 (2.9)	14.5 (10.3)

The results of the citation analysis based on the five subjects are mixed. In the case of the books on "Literature" and "Sociology", neither accessibility nor language are statistically significant. Language is a significant factor for "Public Administration and Political Science" and "Motion Pictures". Only for "History", open access was a statistically relevant factor.

The results of the tweet analysis based on the five subjects follow a different pattern compared to citations. Here, neither accessibility nor language were statistically significant for "Public Administration and Political Science", "Literature" and "Motion Pictures". Language is a significant factor for "Sociology", and – as is the case with citations – accessibility is significant for "History".

Taking these results into account, the conclusion must be that open access does not affect significantly the number of tweets relating to a specific title. However, the influence of language is also limited. Again, the data set is available at http://dx.doi.org/10.17026/dans-x6m-67b2.

9.4.3 Correlating citations and tweets

The background section of this paper describes work by Priem *et al.* (2012) and by Costas *et al.* (2014), in which correlation analysis was used to test for a connection between the number of citations associated with journal articles and altmetrics activity. Of course, correlation is not causation and a connection between citations and altmetrics does not imply such a simple

relation. However, a strong correlation may suggest an underlying cause. Both papers reported a weak but positive correlation between citations and altmetrics activity. In other words, when citations are higher, there is a small chance that altmetrics activity will also be higher.

Haustein *et al.* (2014) link Twitter to a general audience. The altmetrics data in this paper consists of Twitter data, and we might expect only a weak correlation between citations and tweets. In other words: the measured usage of scholarly output – for which the number of citations is used as a proxy – might differ considerably from the interest expressed by the general public – for which the number of tweets is used as a proxy. Lastly, *Figure 3* shows the differences in mean citations and tweets for books with the same subject. This also is an indication of a weak correlation.

A Spearman's correlation has been computed to determine the relationship between the number of citations and tweets in the data set. There was a moderate, positive correlation between citations and tweets ($rs = .299$, $n = 400$, $p < .001$). While keeping in mind the uncertainties described by Thelwall (2016), this result is consistent with the idea that there is not much overlap between academic usage and the interest of a general audience.

9.5 Conclusions

The 2009 monograph experiment was set up to measure the influence of open access using several indicators. During the nine months the experiment ran, it became clear that discovering and consulting the books online benefits strongly from open access. According to the literature on journal article citations and open licenses, a positive influence on monograph citations should have been expected. However, the effect did not occur in 2009. Five years later, the freely accessible books had been cited more on average compared to the control group; a result that confirms the hypothesis that open access has an effect on citations. Yet, when statistical analyses are deployed, the results are more nuanced: when differences in language and subject were controlled for, a small positive effect of OA publishing on citation scores remained.

One of the propositions of making scholarly documents freely available is that it widens access, including for academics who would otherwise not be able to read them. From this follows the assumption that more academic readers will eventually cite the document in their own work. The 2009 experiment demonstrated that online usage benefits from open access, but this usage did not result in more citations. Measuring citations five years

later allowed for the longer time period associated with writing books, which are still a major publication form in the humanities and social sciences. The number of citations measured in 2014 revealed a slight citation advantage for open access books.

A possible explanation can be found in the results of the OAPEN-UK survey of British scholars. Most of the respondents declared that they had little trouble in accessing relevant books, either by borrowing or buying them. Here at least is no indication of diminished access to monographs. As the most likely readers of Dutch language monographs, scholars in the Netherlands and Belgium might work under comparable conditions with relative easy access to academic libraries or funds to purchase books. If that is the case, the significance of free access to online books becomes smaller, although open access might still enhance access.

This study found a similar relationship between open access, subject and language on altmetrics activity associated with books. In the case of OA monographs, making them freely available had a clear positive effect on usage: the free books were used more when compared to a control group of books that are not available in open access. This higher usage has translated into a higher uptake in social media, although the effects of subject and language again played an important role. However, the higher uptake for freely accessible books is not statistically significant.

The results identified very little overlap between Twitter usage and citation behaviour; it seems reasonable to hypothesise that the factors affecting citations of books do not play a significant role in tweets about books. Therefore, the probable reason that open access is a significant influence on book citations does not necessarily apply to Twitter mentions. Nonetheless, it is possible to conclude that making books freely available has some positive impact on the number of tweets. Lowering the access barrier does indeed lead to more attention, in line with the effects for discoverability and online consultation found in the 2009 experiment.

The results also point to the fact that barriers to access are not the only reason for lack of attention. Within the formalized realm of scholarly discourse, the mean number of citations tends to be closely connected to the scholarly discipline. The mean number of tweets per discipline does not follow the same pattern, but there are certainly subjects which are more popular than others. Books on literature, motion pictures or history of art receive a higher number of tweets on average, compared to subjects like history, sociology or law. In other words, the impact of subject should be filtered out, before the effect of open access can be measured.

Apart from subject, language is another factor. This plays a large role in determining the number of both citations and tweets. Publishing books in languages other than English does not only affect usage by scholars, but also the uptake on Twitter. The latter is easily explained by the current preference for English as lingua franca, but also by the fact that scholars are less likely to give attention to other languages. Whether this result is specific to this data set only – because it includes a large portion of Dutch language books – or whether the same result would be found in collections containing a different mix of languages remains to be seen.

9.6 Further investigation: beyond the OA citation adsvantage?

This paper attempts to shed light on the effects of open access on books, rather than on journal articles. The paper has identified that the effect of OA is not as profound for books as it appears to be for journal articles, and that further examination of differences between books and journal articles is warranted. An area of particular interest for future research is whether the slower publication cycle associated with books changes the effect of open access, or whether other factors such as disciplinary culture are responsible for apparent differences.

Another way of looking at the results might be that the OA citation advantage exists, both for articles and books. This has been demonstrated in the case of journal articles again and again. However, more research on the effects of OA on monographs would be welcome, as the amount of published research in this area is small. Still, more interesting questions can be asked.

For instance, if open access helps to disseminate scholarly publications beyond the more affluent academic organisations, will citations and altmetrics reflect this? In other words: will freely available publications be cited more often by scholars working in less privileged circumstances? Or does open access only favour those who would have access anyway? This question could also be investigated through the lens of altmetrics, with a view to establishing whether or not the altmetrics indicators measured are associated with a wider, more global audience.

Earlier in the paper the connection between citation and altmetrics behaviour was discussed. While directly interrogating the reasons for citations or online activities is a complex challenge, this is also an important direction for future research. Understanding whether there are differences between the ways in which research communities perceive OA documents

when compared to closed equivalents may shed light on differences in altmetrics and citation profiles.

Lastly, if the importance of bibliometric analysis as a proxy for research quality is growing, it is vital to understand if there are significant dissimilarities between articles and monographs. Identifying specific differences between journal articles and books and the factors that underlie these differences will enable a comparison of scholarly impact of monographs and articles based on sound principles.

9.7 Limitations

For the purposes of this study tweets referring to book titles were identified through the altmetrics search engine – Topsy.com. The limitations of the Topsy search engine are not known. Furthermore, searching for a book's title may be an imperfect way to find all mentions, due to a lack of online identifiers for monographs.

The method used to collect tweets was geared towards quantitative results: apart from removing tweets on subjects other than the book in question, no attempt was made to analyse the content of individual tweets. Whether authors actively participated in the promotion of their books via social media is not known. However, the author of this paper was employed at Amsterdam University Press from 2007-2014. During that period, no formal policy existed for promoting publications by authors using social media.

Within the analysis, factors other than language and subject were not been corrected for. For instance, the role of document length or publisher's prestige were not accounted for.

9.8 Acknowledgements

The author would like to thank Professor Paul Wouters of the Centre for Science and Technology Studies (CWTS), Professor Frank Huysmans of University of Amsterdam and Associate Professor Lucy Montgomery of Curtin University for commenting on the draft versions of this paper.

10 Conclusions

10.1 Introduction

The deliverance of open access monographs is a complex process. It is based on the actions of many different stakeholders, who have invested time, money and other resources in order to make academic books freely available online. The monographs are disseminated through several platforms, which are part of a larger online ecosystem containing search engines, library catalogues, social media platforms and many more components. Potentially, everyone who is connected to the internet can access the books. The new technical possibilities enabled publishers and funders to offer online collections, while empowering librarians and authors to publish books.

In this publication, the usage of open access monographs is seen as an indication of success. However, the actions of the stakeholders, the complexities of the online ecosystem and properties of the monographs could all affect the usage. To simplify the discussion, I will first discuss the properties of academic books that are not related to online dissemination. In this way, it is possible to make a distinction between aspects that are tied to the concept of the book – whether published in digital or paper form – and the aspects that are connected to the virtual realm. For instance, language is an aspect of the book that affects the usage on online platforms: books in other languages than English are mostly used by native speakers.

I assume that the content of the monograph is created independently of its appearance: the paper version of the book contains the same information as the online version.[1] This assumption has an important consequence: the changes brought on by open access consist largely of adjustments in the online infrastructure. The development of new platforms such as institutional repositories, Google Books, the OAPEN Library etc. are examples of such changes. The performance of these platforms depends partly on their technical specifications, partly on the books themselves.

In the majority of cases, up until the moment the book is ready for publication, the workflow is still firmly grounded in the traditions of the print era; see for example Springer's workflow (Springer, 2017). For some

1 There are exceptions to this rule. An example is the book "Vincent van Gogh – The Letters" and the accompanying website http://vangoghletters.org.

authors, the new possibilities offered by digital publishing could be used to critically engage with the current publication model. Alonso *et al.* (2003) discuss abandoning the print model in favour of the digital possibilities. Hall (2013) is going even further by questioning the concept of the book itself. Instead of a bound entity, the 'new' book is in constant flux: updated through the engagement of researchers and others. At this moment, most books – digital or not – are still a far cry from this vision: a stable text-based publication, consisting of chapters and pages (Carmody, 2011). For now, the current online publication form is basically a digitized version of the paper copy; the same holds true for most journal articles, which also did not change in a significant way (Ware & Mabe, 2015).

Throughout this publication, books are considered to be stable objects, which are not inherently changed by open access dissemination. Several aspects of the book, however, will affect online dissemination. Two aspects have been examined in the previous chapters: language and subject. A third aspect is more implicit: quality in connection to trust. In this concluding chapter I will look in more detail at quality and trust, after a short discussion of language and subject.

The influence of language on dissemination is profound. An author who wants to reach a global audience needs to publish in English. Any other language than English will mainly attract a "local" audience, such as Germany, Austria and Switzerland in the case of German language books, or the Netherlands and Belgium when Dutch language books are on offer. This can be inferred from chapters 6 and 7. This bias towards English also extends to citation indexes, a topic that will be discussed in section 10.5.

In this publication, the subject matter of a book is used as a proxy for scholarly discipline. The scholarly discipline's influence can be found in two areas: dissemination and assessment. In short, subject defines the audience: most of those who are interested in film and media studies are not trying to acquire expertise in the field of archaeology. Bibliometric methods such as citation counts can be seen as a form of assessment. For the humanities and social sciences, this is not without problems (Nederhof, 2006; Ochsner *et al.*, 2017). For instance, each scholarly discipline has different citation practices; which is visible in section 10.5.2. In addition, the results of chapter 9 seem to suggest differences in Twitter mentions per scholarly discipline.

Scholarly research is diverse. Ochsner *et al.* (2017) provide a useful summary of common characteristics. Research in the humanities and social sciences might attempt to accumulate knowledge in the same linear fashion as the natural sciences, or it might be focused on interpreting and reflecting on existing phenomena, such as texts and theories in the humanities and

concepts in the social sciences. In other words, instead of striving towards one definitive answer, it tries to create new perspectives and thus works with competing visions. The researched phenomena can be local, such as the history of a specific region. This also leads to the use of 'local' languages, instead of English.

The diversity in research practices is also reflected in citation culture and quality assessment: each scholarly discipline has different norms whether a publication has sufficient quality. However, consensus exists about one aspect: a publication's quality should not be solely determined by the author. Sufficient quality can only be determined by the author's peers. Both the Royal Netherlands Academy of Arts and Sciences (2010) and Ochsner *et al.* (2012) conclude that quality is best assessed by researchers in the same field, assisted by additional indicators. One of those indicators could be the prestige of the publisher of monographs, as discussed by Giménez-Toledo & Román-Román (2009).

The notion of research quality is determined to a large extent by the scholars within a discipline. Additionally, funders and publishers play a part as well. Funding agencies can influence the research agenda by deciding which research – or scholar – receives subsidy. Moreover, when the role of funding agencies also encompasses publishing open access content – for instance by demanding an open publication license or by using an open access platform – they directly shape the publishing landscape. Publishers play a similar role by deciding whether to accept a manuscript, and by enabling dissemination through open access channels.

Related to the quality of the publications is the issue of trust. Most readers and the libraries and aggregators that act on their behalf will validate the online books on offer. Do they have confidence in the book and the platform it resides on? For instance, when an author publishes a monograph on a personal website, will it reach the same level of usage compared to the same book published on the publisher's website? Intermediaries such as academic libraries might place more trust in the publisher's offering (Moghaddam & Moballeghi, 2007). Another aspect of specialised platforms is their optimisation towards online usage. Not just search engine optimisation, but also by offering services to the intended audiences.

In conclusion, quality assessments directly affect the dissemination of open access books, through the combined actions of the stakeholders. Some groups may act as gatekeepers, strongly affecting the diffusion of books. This is illustrated in chapter 5, where listing titles in the Directory of Open Access Books (DOAB) enhances usage. Before the launch of DOAB, all titles in the OAPEN Library saw comparable levels of usage. When a

set of titles were listed in DOAB – purely based on the licence and not on the contents or the quality of the books – their usage soared compared to the unlisted titles. It seems more than likely that intermediaries accepted DOAB as a valid source, resulting in additional exposure. For instance, the growth of DOAB is listed in the "Dramatic growth of open access" blog (Morisson, 2016). This is further illustrated by McCollough (2017), who sees the Directory of Open Access Books as a tool for discovering open access monographs in academic libraries.

In a sense, the dissemination of books is the final phase of the publication cycle. However, publishing monographs is financially challenging and in section 10.3 I discuss my research on the economic sustainability of a specific model: hybrid publishing. Open access books are disseminated using several platforms, and section 10.4 discusses the optimisation of the infrastructure. After the books have been disseminated, the question arises how to evaluate the results. My answers can be found in section 10.5.

10.2 Web based data sets and data providers

In this section, I will briefly discuss some properties of the analysed data sets. With the exception of the data set of chapter 3, the data have been selected using the web. Collecting data in a web environment is almost by definition automated, eliminating manual procedures and enabling the creation of large data sets in a relative effortless way. However, it also poses challenges. As the environment changes constantly, the gathered data is strongly connected to a certain period in time. For example, the estimated number of websites in 2010 was 200 million, in 2017 the number grew to over 1.7 billion ("Total number of Websites - Internet Live Stats," n.d.). This is also true for the OAPEN Library itself: the number of titles grew from slightly over 850 titles in 2011 to 2,300 books in 2014. In July 2018, the collection comprises almost 5,500 titles. Not just the number of titles increased, also the number of users and the number of book downloads, leading to possible changes in interaction: changes in user's countries; changes in providers and aggregators; changes in the collection's subjects and languages.

On top of this, online tools change or disappear. The data gathered for chapter 8 is based on geographical data provided by Google Books; since 2012, this platform has stopped offering this type of data. The research on monograph citations of chapter 9 is based on Google Scholar. In the year after that research was completed, Google Scholar decided to index the contents of the OAPEN Library (Pinter, 2015). Whether this affects the

number of citations found, is not known. The same chapter also used the services of the Topsy search engine to find tweets. The Topsy.com service has been discontinued in December 2015. Thus, replicating the research on a later date is hardly possible. This is a known problem that affects all researchers working with web based data.

The data sets can be divided in three groups. The data of chapter 3 consists of the sales data of all books published under the same imprint by Amsterdam University Press. The data of chapter 4 to 7 is based on the logged usage data of the OAPEN Library, combined with the metadata describing the books and – where applicable – the added metadata describing the providers. Here, the selection of books is based on all books that were part of the collection during the period under scrutiny. In contrast, the data sets of chapter 8 and 9 are based on a curated and much smaller selection of books. Here, the experimental set and the control set are chosen carefully to remove bias. The data to be analysed is derived from web based platforms: Google Books, Google Scholar and Topsy.com.

When the size of the data sets is compared to the sets used in bibliometric research – for instance in Costas *et al.* (2014); Thelwall *et al.* (2013) – the number of titles is small. Also, the data has mostly been derived from one platform: the OAPEN Library. This might lead to a certain amount of bias. Yet, even the smallest data set is based on nearly 200 books, which are selected carefully to *remove* bias. The larger data sets are based on hundreds of titles, published by dozens of publishers, spanning multiple subjects and several languages. On top of that, the influence of language and subject is analysed separately from the possible effects of open access. Comparing usage data from other platforms would be a good way to enhance our understanding, but comes with its own challenges: differences in infrastructure, collection or definitions of usage must all be accounted for.

I have applied several analytical techniques to the different datasets. In numerous occasions, I applied analysis of variance (ANOVA) to determine whether the influence of one of more aspects of the books or the book's users is more than coincidental. In order to produce reliable outcomes, the values in the data set must be distributed normally; the so-called "bell curve". When the values are out of kilter compared to a normal distribution – which is the case in chapter 9 – I have used the generalised linear model (GLM). The most recent research – described in chapter 6 – was based on social network analysis, combined with a clustering algorithm. The conclusions derived from these analyses will be discussed in more detail in the next sections.

10.3 Economic sustainability

Books are the result of a network of organisations and individuals working together. This network has to be economically sustainable. However, the economic sustainability of monographs has been problematic for decades, long before the advent of open access book publishing. The introduction of chapter 3 describes falling print runs, declining sales and shrinking budgets in academic libraries.

When financing books in a commercial setting is far from easy, how are the costs met if the books are made available for free? What business model can be applied? Publishing monographs in open access could be seen as a "system break" (Pochoda, 2013) or a transition from print-only to digital – mostly in combination with printed books. According to Adema & Ferwerda (2014), this opens new possibilities: increased dissemination, combined with new possibilities to search the contents of collections of books. However, in order to reach this state of affairs it is necessary to find a business model that works.

Several business models for academic books have been discussed in the literature. Greco & Wharton (2008) recommend looking into a model optimized for open access books, combined with a print on demand system, for those who still prefer a paper version of the book. The search for new business models is also described by Withey *et al.* (2011), who are investigating how to preserve the best elements of the current publishing system in a new era of open access monographs. There are numerous other business models, ranging from a hybrid publication model to crowd-funding (Ferwerda, 2014). Recently, Knöchelmann (2017) discussed the open access book market, tying successful upscaling to funding.

Within the direct sphere of influence of the OAPEN Library and the Directory of Open Access Books (DOAB), different business models are used. For instance, the French organisation OpenEdition – which makes titles available via DOAB – offers a "freemium" package to libraries: a combination of a basic version of a publication that is freely available online, combined with paid-for premium services (Mounier, 2011). Knowledge Unlatched is another example, using a model based on the cooperation with university libraries. It establishes a library consortium that pays a "Title Fee" to a group of publishers. In return, the publishers offer print copies to member libraries at a discount and also make the books available in open access. One of the deployed platforms is the OAPEN Library (Pinter, 2012).

Very few papers can be found on the costs of producing monographs. The costs of creating a – paper only – monograph by an American university

press is discussed by Wasserman (1998). Roughly speaking, the costs of around $24,000 are not met by the expected sales: depending on the edition, the losses range between $8,000 and $13,000. A more recent investigation by Ferwerda *et al.* (2013) looked into the publication costs of Dutch monographs publishing. Based on the budgets of 50 books the average costs for publishing a monograph in the Netherlands was found to be slightly over €12,000. And finally, Maron *et al.* (2016) examined the publication costs of 20 American publishers in 2014. The average costs of a digital monograph ranged from $30,000 to $49,000. Whether each amount is based on the same cost structure is unknown. Recently, Pinter (2018) explains that comparing monograph publications costs is problematic, due to the diversity of publishers.

It is doubtful whether the economic sustainability of monographs is guaranteed by the sales of paper copies, and the literature on costs seems to suggest that a substantial amount of money is needed to produce an academic book. In such circumstances, will open access publishing have a positive monetary effect?

Chapter 2 offered some further insight into the economic effects of open access monograph publishing, by examining the effects of a hybrid business model. In such a model, paper copies of books are sold, while an online version is also made available for free. The main assumption is that the open access version of the title acts as an "advertisement": when the reader has discovered the book online, this will possibly lead to the purchase of the paper version, as many readers still have a strong preference for the paper codex. As a counterargument, one might argue that paper books are not a necessity in the era of e-book readers and high-quality tablets. The main question is thus whether the hybrid business model enhances or diminishes sales.

The data underlying chapter 3 does not come from a controlled environment, such as described in Snijder (2010). Instead, it examines sales data obtained from a "normal" business setting: sales data from Amsterdam University Press obtained in the period 2010 to 2012. While the publisher uses the hybrid business model – selling paper copies alongside online open access versions – the commercial expectations for the open access titles differ from the closed access titles. This can be inferred from the print run: a higher print run indicates a higher expectation of number of copies sold. The average print run of books published in closed access was much higher, compared to the open access titles. Apart from commercial potential, the moment of sales is also an important factor: most copies are sold in the first year of publication.

Chapter 3 set out to measure the influence of open access on monograph sales. Furthermore, the effect of open access was compared to other influences on monograph sales: commercial potential, front list and back list, and language. Each influence is statistically relevant, making it harder to single out the effects of open access. The difference in number of copies sold in the first year – the front list – compared to the number of copies sold in the subsequent years – the back list – is striking: the mean sales in the first year is about five times larger than the year after that. Consequently, I analysed the front list sales and the back-list sales separately.

The results of the front list sales can be explained by a combination of commercial potential and language; open access publishing does not have an effect in this situation. The results for the back list are similar to the front list outcomes. The influence of language was not statistically relevant, and open access publishing is a relevant influence on sales in certain cases only: the subsets of books whose print run is between 1 and 2000. The resulting average number of copies sold seems to point to a small advantage for the closed access titles. Whether the advantage of closed access published books is economically relevant, is questionable. Over 65% of all copies sold were open access titles.

In the debate on the economic sustainability, the small differences in the number of copies sold are not the main issue. In all discussed experiments open access did not have a large effect on monograph sales, positive nor negative. At the start of this section, I mentioned the problems in the book trade, and I have found that the hybrid model does not lead to more sales.

10.4 Factors affecting dissemination

So far, I have discussed the aspects of the books which remain stable in a paper and a digital environment and the financial fundament under monograph publishing. The next aspect to explore is online dissemination. The distribution of open access monographs consists of two parts: a digital collection and the means of dissemination. In the previous chapters, several platforms were introduced: institutional repositories, publisher's collections, the Google Books platform, the OAPEN Library and the Directory of Open Access Books (DOAB). Some aspects of this non-exhaustive list will be summarized in this section, as an illustration of the open access monographs infrastructure.

Each platform has its own affordances. For instance, disseminating books via an institutional repository may underline the relation with the hosting

organisation. The Google Books platform enables different things: besides being directly linked to the Google search engine, it allows rights owners precise control over how much of the book is made visible to the public. A platform such as the OAPEN Library is optimized for disseminating OA books via several channels. The Directory of Open Access Books only stores metadata, but amplifies the use of the titles listed.

Understanding the strengths and weaknesses of platforms is vital for choosing a dissemination strategy. Online dissemination platforms shape what the readers can do with the book, which affects its usage. The usage data generated by the platform can be used to assess the impact of the books on the platform, an idea investigated by Herb *et al.* (2010) and in the previous chapters. What a platform is capable of, is decided by its owner. Each owner will have different preferences, leading to a landscape of various possibilities. To illuminate the differences, I will shortly discuss the platforms and their owners.

Institutional repositories are based on a set of standards promoting interoperability. Each repository should be able to connect to other repositories and use its content. They could be seen as a natural extension of academic libraries: in most cases the library will manage the repository. Other platforms are also used within the library community: some librarians make a part of their collection searchable through Google Books ("Library Partners – Google Books," n.d.). Platforms such as the OAPEN Library or DOAB are also used as a source for OA books. Apart from academic libraries, some funding institutions may choose to directly deploy repositories or comparable platforms. For instance, the Austrian science fund FWF directly places books in the OAPEN Library (Snijder, 2015). Others, for instance the Spanish National Research Council, have chosen to set up an institutional repository (Bernal, 2013).

Some publishers – for example Brill or ANU Press – have made a collection of books available on their website. Setting up a bespoke platform enables publishers to control what data to collect about the users. Some people will argue that knowing more about the people active on a platform solely benefits the platform owner. A recent example is the speculation by Kelty (2016) about the motives of Elsevier to purchase the SSRN platform: SSRN's data can be used as a means to evaluate scholars; to be sold to university administrators. The question of privacy is discussed further in chapter 6: how to balance the privacy of the readers versus the desire to know the "customer" in detail? A publisher might also use other platforms to distribute open access books online: Google Books, OAPEN Library or DOAB.

Strictly speaking, the Google Books platform is not an open access platform. It is a search engine that contains and indexes books, which also allows the rights owner to decide how much of the book's contents is publicly visible. This feature enables publishers to fully open up a book if desired. Controlling the visibility of the book's content can be used to set up experiments in which a set of books with a limited amount of visible content is compared to a collection of books where all pages were visible (Snijder, 2010). However, publishers and libraries do not control the platform, and the platform's owners decisions may not always suit them. For instance, since 2015 no new publishers are allowed to sign up to the Google Book platform (S. Hall, 2016).

The platforms differ in capabilities, but also in content. Each platform strives to maximise its use – at least within its target audience – and a major factor is the quality of the offering. Thus, I assume that each platform will select suitable titles and refuse inappropriate ones. What is a suitable collection will be different for each platform: institutional repositories and publisher's platforms will be limited to their organisations; the OAPEN Library and DOAB collect titles from different publishers but emphasize quality assurance of the titles; the Google Books platform attempts to keep pirated books from their collection.

Maintaining a trusted platform might also be a strategic advantage for the hosting organisations. For publishers, it may be a way to directly sell copies to readers – cutting out the middle man. For academic libraries, it may be a way to strengthen their position within the university, and a possible counterweight to the influence of publishers. In the case of large commercial organisations, the platform may be part of other offerings. The success of Google depends at least partly on knowing the preferences of their users. The kind of information gathered may lead to privacy concerns. This conflict of interests has been discussed in more detail in chapter 6.

Open access book platforms are still a relative new phenomenon. What aspects are important for the dissemination of open access books? In the next section, I will discuss several of these aspects.

10.4.1 What works in digital dissemination?

The research in the previous chapters is based on experiments, carried out on several platforms. Most experiments have taken place using the OAPEN Library. The OAPEN Library has been operational since 2010, making it one of the longest running open access monograph platforms. It has several properties that help examine the influences on the usage of

monographs. Firstly, its collection of several thousand books contains large groups of books in several languages, especially English, German, Dutch and Italian. Furthermore, the collection spans a broad range of subjects. The monographs are not only available through the OAPEN interface, but – through availability of metadata and agreements with commercial and non-commercial aggregators – are also directly accessible via library catalogues and other platforms. Due to the fact that the platform has been operational for several years, trends over longer periods can be examined. The diversity in licences is another factor that can be studied. Lastly, the books made available on this platform have been vetted through a peer review process.

Before, I have discussed economic sustainability as a basic requirement for disseminating open access books. Now I will look into another aspect affecting the distribution of open access monographs: dissemination channels. Online dissemination contains more than placing documents on a website, hoping they will magically turn up prominently in the results of search engines. Instead, it is necessary to use the channels that are best suited for the targeted audience. Until recently, in the literature on open access, dissemination channels seem to be a given. If it is discussed at all, dissemination is described as making papers available in an institutional repository.

In chapter 4, the success rate of two dissemination modes has been examined: the OAPEN website acting as an Online Public Access Catalogue (OPAC), and direct access where the reader directly downloads the book without searching the website. A "direct" download implies that the reader has used other means to find and select the book. The direct search channel is based on metadata only, which is incorporated into systems outside the OAPEN Library. The usage data obtained comes from three channels: through the website only; a combination of website and direct downloads; or downloads only. The data is analysed both quantitatively and qualitatively in chapter 4. The quantitative analysis reveals a large difference between the number of books that were downloaded without searching the OAPEN website and the other dissemination channels: 73 % of all downloads can be attributed to 'direct' downloads. The results of the qualitative analysis are not so easy to interpret: the provider's characteristics nor the properties of the books were statistically significant.

The books were downloaded through providers, which I categorised in two ways: the type of provider and the state of their country's internet infrastructure. This categorisation was introduced in chapter 7, which will be discussed later. The question is whether a connection between

the provider category and dissemination channel exists. Regardless of the channel, most of the usage comes from three types of provider: academic, internet service providers (ISP's) and ISP's from a country with a highly developed internet infrastructure.

The state of development of a country's internet infrastructure does not affect which dissemination channel is used. However, the digital divide is clearly visible in the smaller usage from the countries with a less-developed internet infrastructure compared to the small group of better equipped countries. Lastly, the subject or language of the downloaded books did not affect the usage of the channels.

A possible explanation for the large percentage of direct downloads can be found in the theoretical models on the use of innovations. Whether or not a new system is used depends on several aspects, such as its fit with existing usage patterns, perceived ease of use and social norms. It is possible that most users prefer their 'own' systems, with which they are familiar and which are part of their routine and environment. In that case, learning to use a new interface may not be seen as a worthwhile investment.

The high percentage of direct downloads – over 70% of all book downloads – cannot be fully explained by search engine optimisation, as only 30% of the internet traffic to the OAPEN Library during that period originated from search engines. This means that a sizable portion originated from other types of websites. The only way to directly download the books is by using a specific download address. Those addresses are distributed by the OAPEN Library, through its metadata feeds. When other systems or websites incorporate the web addresses that enable direct downloads of books, they act as aggregators. While I did not examine this, it is likely that some websites only display a portion of the collection. An example is the Ancient World Online blog (Jones, n.d.), which lists only monographs about the Antiquities Period.

Before, I stated that the success of open access publishing depends on many stakeholders. The main purpose of open access is to make knowledge available, and it is useful to investigate the factors that enhance dissemination. The results of chapter 4 reveal an important aspect of open access dissemination: enabling incorporation into other systems enhances the monograph usage. Here, the solution offered by the OAPEN Library is providing metadata to be used by aggregators. While the metadata is available to all, a relative small portion of usage can be attributed to search engines. The indexation by search engines is an automated process, but the incorporation of the metadata into other systems – which aggregate information for readers – is the result of a conscious decision. I conclude

that this decision is based on trust. Aggregators accept the monographs offered by the OAPEN Library as a viable source, and make them available to their patrons.

The importance of aggregators is also visible in the results of chapter 5: their influence on usage is much stronger than that of licenses. Within the literature on open access, the role of licenses is discussed extensively. According to the Open Access Scholarly Publishers Association (OASPA), "true" open access can only be achieved through the use of a specific Creative Commons license: CC-BY (Redhead, 2012). If true open access means optimal dissemination of scholarly content, books published under an open license – which allows sharing it contents – should perform better than books made available under a license that permits nothing more than downloading for personal use. I tested this hypothesis on the OAPEN Library, where roughly half of the collection is available under a license permitting reuse, and the other half under a license that only permits personal use. The results showed that the number of downloads of open licensed books did not differ significantly from the monographs with a "free to read" license.

However, I also investigated the role of the Directory of Open Access Books (DOAB), by examining the usage data of the same collection after the launch of DOAB. The DOAB aggregates open access books, but only those with an open license. Open licenses such as Creative Commons are machine-readable: they can be used in automated processes, leading to new possibilities. In the case of the OAPEN Library, the licensing information is part of the metadata. The metadata is used by the DOAB, in order to select books with an open license. When the period after the launch of DOAB was examined, the difference is far greater. Books listed in the DOAB have been downloaded almost twice as much on average compared to the other group of titles. Even when allowing for the role of subject and language, the influence of DOAB is profound.

While the license is seen by many in the scholarly communication field as an important enabler for open access, it is doubtful whether the readers care as much. The results seem to suggest that a "free to share" license is not an important incentive compared to a "free to read" license. The number of downloads was not boosted by an open license, the usage was boosted by incorporation of a new service: DOAB. It is DOAB policy to only list monographs with an open license, and thus half of the OAPEN Library collection was imported, leading to the large difference in usage.

The influence of other aggregators could explain the large uptake of the books listed in DOAB. When more aggregators are aware of the existence of DOAB, compared to the OAPEN Library, the monographs listed in DOAB

will receive more attention. Several authors see DOAB as a comprehensive source of open access monographs (McCollough, 2017; Morisson, 2016). More exposure will also lead to more data usage. Again, usage is strongly affected by trust: being listed in DOAB – a widely trusted source – results in more aggregation and thus more visibility, which stimulates the usage of open access monographs.

As far as usage by readers is concerned, the results of chapter 5 seem to downplay the role of licenses. Given the fact that a large percentage of the books were published under a CC-BY-NC-ND license – which does not permit commercial use or creation of derivative works – other stakeholders might consider those licenses as equal to 'free to read' licenses. For instance, research institutes may be more strongly bound to the terms of the licenses, especially when a large set of books is examined. The use of large corpora for text mining depends on permissions by rights holders (Van Noorden, 2014). Still, the influence of aggregation in DOAB is undeniable: even the books published under the most restricted open licence have been used more, compared to the books available for reading purposes only.

In conclusion, while licenses are important for certain groups of users, this is not the case for those who want to read the books. For them, usage is not boosted by licenses, but by the choices of aggregators.

10.4.2 Clustering books and readers

So far, I have looked at book dissemination purely based on numbers; examining factors affecting the number of downloads, a proxy for the number of times a book is read. Chapter 6 uses a different angle: creating clusters of books that are suitable for a group of readers. Instead of lumping the users of the OAPEN Library together into large groups such as academics, government employees or the general public, an attempt is made to uncover "communities": groups of people that share an interest. Defining communities and finding suitable titles is an important task of libraries. Online retailers such as Amazon use a different strategy, based on personal recommendations. Creating a more fine-grained understanding of the users of any open access platform helps to deliver the best titles. However, it also leads to questions of privacy: is it desirable to store information about individuals? These questions are examined in chapter 6.

One of the most prominent success factors of online retailers is the amount of knowledge they possess about their customers. If the preferences of each client are known, it is possible to offer desirable products. In such circumstances, the online retailer will strive to maximise the amount

of known facts about all their customers. I noted before that collecting and storing data about individuals leads to discussions about privacy. For libraries, the protection of their patron's privacy is an important part of their core values. Online dissemination platforms could model themselves after online retailers; after all, apart from charging money for their services, they perform more or less the same functions. However, the main purpose of open access platforms is not to maximise sales, but to maximise the usage of documents, which is closer to the core values of libraries.

I investigated whether it is possible to create optimized recommendations while storing a minimum amount of information about individuals. A solution for this problem might be found in the download behaviour of all users of a dissemination platform. By analysing all data at once, instead of focusing on individuals, it might be possible to discern patterns: clusters of related books that are downloaded together. If such clusters can be found, they could form the basis of a recommendation, akin to recommendations by online retailers. To create an optimal solution, it is also necessary to understand who is interested in a specific cluster of books, without targeting individuals. To resolve this, the research focuses on finding communities: groups that share a common trait.

The research was based on two data sets, consisting of providers, books and the number of times a book was downloaded. The first set was captured during 2012 and the next set is based on data from 2014. Each book in the collection was categorised through its language and subject. The information about providers is limited to name and country of origin. The linked titles and providers are clustered using the Wakita-Tsurumi (2007) algorithm, resulting in dozens of clusters. The ten biggest clusters were analysed, comparing the books' language and subject and the providers' nationalities to the complete data set.

Within the examined data, several clusters could be identified that were not the result of random downloads. Some clusters contain large percentages of non-English books, combined with a large set of providers consisting of native speakers. An example is a cluster containing Dutch language books combined with many providers from the Netherlands and Belgium. Other clusters – where the language is mostly English – contain books on certain subject, such as film and media studies or Indonesia and South-East Asia. When the subject is region-based, this is also reflected in the nationality of the providers.

The clusters are not created manually, but are the result of an algorithm. Consequently, this procedure can be part of an automated process, akin to the recommendation services of online retailers but without violating the

privacy of individuals. However, there is still room for improvement: clusters found in 2012 are not visible in the 2014 data. This is not uncommon: other research on clustering techniques also show differences, all of which might be valid in their own right (Gläser *et al.*, 2017).

In conclusion, to a certain extent it is possible to use clustering algorithms to create optimized recommendations, while still protecting the privacy of individual readers. Optimized recommendations by open access platforms should lead to higher usage of open access monographs. The results of chapter 6 can be seen as a proof of concept, to be further refined.

10.5 Evaluation of results

Until now, I discussed the hybrid business model and several aspects of digital dissemination affecting the usage of open access monographs. From these practical considerations I will now move to the outcomes: does publishing monographs in open access lead to a greater scholarly impact and societal influence? To answer this question, I first need to define scholarly and societal impact, insofar as it applies to monographs. Open access monographs can have an impact on the work of academics – I will categorize this as academic or scholarly impact – and they might affect those who do not have access to large academic libraries – defined here as social or societal impact.

Monographs require other indicators than journal articles. Bibliometric measurements like the journal impact factor have been used for decades (Garfield, 2006). For monographs, similar data is not abundantly available; instead, metrics based on library holdings might be used. My research is based on usage data, derived from online platforms. In this case, the proxy value for academic impact is the amount of usage originating from academic institutions, compared to usage from other organisations. This metric is restricted to the number of academics who use the internet infrastructure of their institution to access the OAPEN Library; it will not take into account academics who use other internet providers. My research on the academic impact of open access monographs is not limited to usage data: in chapter 9 I have examined whether open access affects the number of citations.

I have examined the social impact of open access monographs using indicators based on usage data. When the usage originates from governmental, non-profit or business organisations, I have classified this as types of social impact. Another indicator of the social impact of monographs can be found in altmetrics, here defined as online activity about academic publications.

Some types of online activity are closely tied to the work of academics, for instance Mendeley.com or ResearchGate.com. Others – such as Facebook or Twitter – are used by a large section of the general public. Mentions of open access monographs on those platforms stand a larger chance to come from non-academics.

In short, in this section three types of indicators will be discussed: citation based indicators, platform usage and altmetrics. The definition and mutual relations of these indicators is discussed in more detail by Glänzel & Gorraiz (2015), who state that the combination of usage, altmetrics and citations leads to a more complete view of a document's impact. According to the authors, citations are an accepted indicator of academic impact, but do not capture social impact. Usage measures the intention to read documents and altmetrics indicate mentions of documents, both in academia and beyond.

10.5.1 Impact measured

Indicators of academic impact are relatively easy to identify through usage originating from academic institutions. Social impact is more diverse: it encompasses usage by non-academic readers with a professional interest such as government employees, but also readers without a professional interest: members of the general public. To distinguish between these groups, I use the connection to an organisation – which can be inferred from the usage data – other than an Internet Service Provider or an academic institution. I assume that non-academic readers with a professional interest are connected to an organisation.

The defining characteristic of members of the general public is their lack of connection to an organisation. This complicates identification based on usage data: if readers use an Internet Service Provider (ISP), does that mean they are not connected to an organisation, does it mean that "their" organisation is unable to provide direct internet access, or are they just not using their organisation's equipment? Differences in internet infrastructure are also at the root of the digital divide between developing and developed countries, leading to the question whether open access leads to more usage when the available internet infrastructure is not optimal.

Categorizing users in groups is useful to distinguish between the usage by academics and usage by others. Simply put: usage of monographs by non-academics is a form of social impact. Comparing the percentage of non-academic users of a set of open access monographs to a set of monographs in closed access helps to determine whether open access leads to a higher

level of social impact. Thus, we are able to test the assumption that open access monographs' availability beyond academic institutions leads to more usage by non-academics. However, other influences may affect usage. Differences in available infrastructure – the digital divide – is an example. Another possible factor is the dissemination platform: is it able to reach non-academics? Furthermore, aspects of the books such as language and subject play an important role. Any conclusion about the social impact of open access monographs based on usage data must account for these factors.

Besides usage data, other indicators of social and academic impact are also available. In the realm of journal articles, the number of citations is the most-used metric to assess academic impact. For monographs, citations are more problematic, which has been discussed in chapter 9. Investigating citation data for books is hampered by a lower availability of indexation services. Another challenging issue is the slower pace of citations, leading to a "citation window" of at least six to eight years. The third factor might be the difference in citation culture between scholarly disciplines. Lastly, in some fields of HSS, writing in English is not always the norm; this is problematic when citation indexes might be biased toward Anglo-Saxon regions (Nederhof, 2006). As is the case with usage data, any conclusion about the academic impact must take into account the special circumstances around open access monographs.

An indication of social impact might be found using altmetrics. Altmetrics share much characteristics with usage data. Instead of counting activities from infrastructure that is directly connected to documents, the usage of a broad range of social media and other online outlets is measured. As is the case with online book platforms, some outlets are more strongly directed towards academic users, while others are more open to everybody. For instance, online reference managers such as Mendeley or specialised websites such as ResearchGate are far more used by academics, while platforms such as Twitter or Facebook have a more diverse user base. On top of this, the different altmetrics outlets are also aligned differently to document types. Hammarfelt (2014) concludes that Mendeley is the best altmetrics outlet for humanities articles, while books are mostly mentioned on Twitter. In conclusion, Twitter is most likely to be used by the general public and mentions books most often. For that reason, the number of tweets is used as an indicator of social impact in chapter 9.

The next section discusses several examinations of the impact of open access on academics and non-academics.

10.5.2 Indications of impact

The question examined in chapter 7 is how to provide quantitative evidence of both academic and social impact of HSS research. The use of bibliometric data for monographs is problematic and the humanities and social sciences tend to place more emphasis on the societal impact of the results. Delivering evidence of impact depends for a large part on either self-reporting or in-depth discussion with stakeholders. Both methods are labour-intensive and susceptible to bias. Here, taking advantage of usage data might help to display another aspect: interaction with published results. Like altmetrics, the usage data is the direct result of online interaction, and the large number of data points enables the creation of sophisticated reports.

The usage data contains information about the organisation through which the reader accesses the web. By determining the type of organisation and the country of origin it is possible to assess the impact of the books, both in academia and beyond. The methods – tested on the OAPEN Library in 2011 – helps to uncover stakeholders, who may not always be known beforehand. Over 27% of the data is directly linked to academic users. In contrast, the usage linked directly to other "professional" users is less than 5%. The remaining 67% cannot be directly ascribed to the general public. The type of provider is a commercial Internet Service Provider (ISP), making it impossible to determine what organisation – if any – the reader is associated with.

In order to better categorize this large group of readers, I combined the available information about the country of origin with the state of intranet infrastructure. By using a fairly strict threshold, countries were grouped in those with a highly developed internet infrastructure, and those without. I assume that readers from a country with a highly developed internet infrastructure who download monographs out of a professional interest are more likely to use their organisation's internet infrastructure instead of an ISP. Thus, readers based in countries with a highly developed internet infrastructure that use an ISP to access the monographs, are more likely to be part of the general public. In this way, the large group of uncategorized users – 67% – can be classified. The smaller half of this group is still not categorizable, but the other half might be part of the general public in the wealthier countries of the world.

Apart from using provider types as proxies for users, the influence of scholarly discipline was analysed by looking at the differences in usage for humanities and social sciences books. Also, the differences in geographical impact of books in English versus books in Dutch are quite visible.

The question I examine is not whether more people are interacting with the monographs, the question is what kind of people are using the open access books. Usage should always be evaluated within the context of the platform. For instance, measuring usage of an academic library will not lead to finding many non-academic readers. The OAPEN Library is freely accessible and has taken several measures to make its content widely used, which might help to attract many different users.

When the results of chapter 4 are considered, we see that three-quarters of usage stems from direct access: incorporation into other systems than the OAPEN Library interface. The usage percentage from academic providers is less than 20%, while the usage through ISPs operating in countries with a highly developed internet infrastructure is 50%. Both the results of chapter 4 and of chapter 7 point to a relative low usage directly linked to academic institutions. The results of chapter 4 seem to suggest that other platforms than the OAPEN Library incorporate descriptions of the books. However, a large percentage of those platforms are not directly linked to an academic institution.

Thus, given these results it is feasible that the OAPEN Library's contents are available to readers beyond academic institutions in the "global north". The percentages directly linked to readers with a professional interest – those linked to government, non-profit or business organisations – are invariably low. And the largest single category consists of internet providers that have – at the very least – a possible link to readers that have downloaded the books for other than professional reasons. Returning to assessment within the context of a platform, its potential reach is wider than academic institutions alone. Consequently, the books available at an open access dissemination platform stand a good chance of reaching a wider audience. The percentage of monographs that are downloaded frequently and by other categories than academic institutions alone, are an indication of social impact.

Social impact is not restricted to the "global north"; does open access help to bridge the digital divide between those living in the richest countries and those in other parts of the world? Chapter 8 surveys whether open access enhances the use in the developing countries. In other words: does open access help to overcome the inequality in access to the internet, both in a technical sense and in lack of knowledge to optimally use the available resources? To test this, the usage data of the books in the experiment performed by Snijder (2010) were combined with geographical user data.

During the experiment – run in 2009 – several sets of monographs were made freely available. Another set of books was used as a control group.

The data was gathered from the Google Books platform; access to the books was strictly controlled for the experiment. This platform was ideally suited for this type of experiments: while all books on the Google Books platform were fully indexed by the Google search engine, it allowed publishers to decide what percentage of the book's contents were freely available. Thus, some books could be fully read online – 100% of the content available – and the control books showed no more than 10% of the text. To remove bias, the sets were carefully set up, based on subject; type of work; expected sales and publication date. The analysis in chapter 8 is based on 180 English language monographs.

Even when using a platform that is part of a globally used search engine, the digital divide between developed and developing countries is clearly visible: only 30% of the usage comes from developing countries. Further analysis of the differences between the usage of the open access books versus the closed access books revealed a more positive outcome. When reviewing the usage from developing countries and developed countries, the relative usage of open access monographs by developing counties was higher compared to the usage of the books that were not completely available. This is an indication of social impact: more usage of open access monographs by those in a disadvantaged position.

Before, I examined the possible influence of the collection's geographical focus and usage from the same region. While it might be a factor contributing to the lower usage from developing countries, the setup of the experimental and the control set of books helped to evenly spread the subjects. Additionally, the differences in internet infrastructure will have played a role in access and – in this case – the positive influence of open access is visible.

So far, I discussed usage as a means to measure academic and social impact of open access monographs. When the users are categorized by organisation type, academic users are the largest group. However, the combined download figures from academic organisations amount to roughly 20% of all downloads. In other words: it is possible to show the academic impact of open access monographs, but based on this data it is hard to conclude that open access enhances usage among academics. When I look at social impact, the results point toward increased usage by those who normally face additional challenges to access scholarly books: non-academics in the "global north" and those living and working in developing countries.

In order to answer the question whether open access has a positive influence in academia, I turned to another measurement: the number of citations. Many open access advocates have discussed the positive influence

on citations – seen by many as a major indicator of academic impact. Can a 'citation advantage' for open access monographs be found? This has been investigated numerous times for journal articles, but scarcely for books.

The research of chapter 9 dealt with these aspects in several ways. Instead of relying on a citation index, I used the Google Scholar platform. Secondly, to account for the "citation window", the examined books were published at least five years before the date of obtaining the data. The differences in disciplines – and languages – have been dealt with in several ways. To maintain a balanced division of subjects and languages, I used the same sets of books as in Snijder (2010). Furthermore, I studied the influence of books in humanities versus other disciplines, plus additional testing on several groups of books on more specialised subjects.

To examine social impact, I used altmetrics. As mentioned before, some altmetrics sources are geared towards academic users, while others target a more diverse audience. Reference managers such as Mendeley are strongly related to academic use, while platforms like Facebook or Twitter are used by all types of internet users. In this research, I selected the altmetrics platform that performs best on monographs and is mostly connected to the general public: Twitter.

Given these preparations and choices, is it possible to establish whether open access has a positive effect on the number of citations? Also, does open access lead to more uptake by the general public? Looking at citations, the results are more or less in line with the literature on journal articles: a small but statistically significant positive effect of open access on the number of citations, even when the analysis takes into account the influence of language and subject. For tweets, the situation is slightly different. In the same way as citations, the average number of tweets about open access monographs is larger than the number of tweets about closed access books. However, the difference is not statistically significant. Lastly, little overlap exists between Twitter usage and citation behaviour. Thus, open access does not affect Twitter mentions in the same way as citations.

If citations are an indication of scholarly impact, the results point to positive influence of open access. However, the influence is not very large. A possible explanation might be found in the fact that a large proportion of researchers who are interested in the books in the data set, are in the position to access its contents anyway. This is supported by the outcomes of in chapter 8: in 2009, over 70% of usage of the English language books – which have a more global audience compared to the Dutch language books – was connected to the richest countries. Presumably, readers working in

those countries have a far better chance to view the books' contents, either through a library or by buying a copy.

Does this mean that the higher uptake of open access books – see Emery *et al.* (2017); Ferwerda *et al.* (2013); Snijder (2010) – can be mainly attributed to non-academics? Given the results of chapter 6, where roughly one-third of the usage has a larger chance to be associated with the general public, this might seem plausible. However, this conclusion is supported by indirect evidence. First of all, the data set of chapter 8 is based on 400 books. Whether this set is large enough to warrant such broad conclusions is questionable. Secondly, if I assume that the Twitter usage in chapter 9 indicates interest by the general public, the lack of statistically significant evidence is problematic. And lastly, I have discussed the differences in platforms. The users of the Google Book platform might differ significantly from the users of the OAPEN Library, and any conclusion spanning multiple platforms should be backed by solid evidence.

Direct evidence of the societal impact of open access monographs beyond the downloads of businesses, governmental organisations and non-profit organisation is not easy to obtain. Likewise, knowing what usage is related to the general public – which is by definition not affiliated to a specific type of organisation – is also problematic. Compared to journal articles, the available research data is still scarce. Therefore, more data is needed to provide more definitive answers, especially usage data and data about the collections of other open access book platforms. This will enable us to compare the effects of the identified factors on platforms with other collections and affordances: what are the effects on usage, citations or altmetrics? Hopefully, my research marks the start of more investigations.

10.6 Concluding remarks: factors affecting usage and the impact of open access

The introduction states that the level of open access monographs usage is primarily determined by book-related factors such as language and scholarly field or the configuration of dissemination platforms. The results show that these factors indeed affect the usage. Another factor is the level of trust in the content on offer. Contrary to expectations from several open access advocates, open licenses do not affect the level of usage. Furthermore, open access does not lead to more sales of monographs, yet it enhances usage in developing countries and the number of citations.

Most experiments in this publication have involved the collection of the OAPEN Library; a diverse set of books spanning multiple disciplines and languages. Therefore, it was relatively easy to measure how subject as proxy for scholarly field and language play an important role. For instance, while the topic of migration is not only discussed in academic circles but also in most newspapers, the audience for Sumerian spells[1] might be smaller. Usage is also connected to the geographical location of the readers: academic books discussing a certain part of the world tend to be read more by those who come from the same region. The usage of monographs written in other languages than English is also affected by geographic factors: books in German are more downloaded in German-speaking countries; the usage of Dutch language books is highest in The Netherlands.

The role of subject and language was to be expected. Furthermore, it is obvious that online dissemination is affected by the infrastructure that supports it. This has been clearly visible in the digital divide between rich and less well-off countries. Another aspect of online dissemination infrastructure is its interconnectivity: how well does one source integrate into another platform? The fact that the majority of the OAPEN Library downloads does not involve the front end can be seen as an illustration of the immersion into other systems.

Whether the technical abilities of dissemination platforms such as the OAPEN Library or the Directory of Open Access Books are used depends on a far less obvious factor: trust. Making a book available online does not automatically lead to optimal usage. Most people rely on filtering mechanisms to separate the wheat from the chaff. These mechanisms may include library catalogues, mentions on social media, specialised websites or blogs and many more possibilities. Additionally, the "filters" may rely on other sources: for instance, libraries might employ content aggregators.

In short, whether an open access monograph – or a platform that disseminates open access monographs – is accepted, depends on a conscious decision, not solely on an automated process. This is illustrated by the added usage from inclusion into the Directory of Open Access Books, but also by the inclusion of the contents of the OAPEN Library into other systems.

Ultimately, the decision to use an open access book platform is based on trust. Trust and the notion of quality are closely connected: when the books on offer are of sufficient quality, the prospective readers – or aggregators – will take action to obtain one or more books. As it is unlikely that each

1 Schramm, W. (2008). *Ein Compendium sumerisch-akkadischer Beschwörungen*. Universitätsverlag Göttingen. Retrieved from http://www.oapen.org/record/610352

book on the open access book platform will be vetted before downloading, the prospective readers – or the aggregators acting on their behalf – must assume that the offering is of sufficient quality. In other words, the readers must put their trust in the choices made by the platform.

Subject, language, infrastructure and trust are all influences that shape the usage of open access monographs. Other factors are not as important: licenses and the effects on sales. Licenses are seen as an important part of open access: the ability for readers to reuse the content has been described explicitly in the BOAI (Chan *et al.*, 2002). Given the emphasis on reuse, it was reasonable to expect more usage of monographs made available under a licence that actually permits it. However, whether an open access monograph licence only permits reading and downloading for personal use or enables content-sharing did not matter. Thus, the influence of licenses on usage is negligible.

The conclusion that open access does not affect the sales of monographs is not very surprising. I have been involved – directly and indirectly – in several experiments to measure the effect of open access on monograph sales (Collins & Milloy, 2016; Ferwerda *et al.*, 2013; Snijder, 2010; SNSF, 2015). In contrast to chapter 3, these experiments are based on a careful selection of monographs: an experimental set of titles that are published in open access, and a control group consisting of comparable books. None of these experiments resulted in a significant increase or decrease of the number of copies sold for the set of open access monographs.

Open access to monographs leads however to more usage in developing countries, a positive result. One of the goals of open access is enhancing the usage by those who would otherwise not be able to read scholarly output. Here, this goal has been achieved, albeit on a small scale. Another often-used benchmark in the realm of journal articles is the "citation advantage" of open access publications. For monographs, I was able to demonstrate a slight citation advantage.

To recapitulate, while open access monographs dissemination is only possible by removing paywalls, the level of usage is primarily determined by language, subject, infrastructure and trust. Given these influences, open access enhances usage in developing countries and the number of citations.

10.7 Practical implications and further research

What are the practical implications of these results? In my opinion, an open access monographs platform should focus on trust. After all, trust is

the most important aspect: any platform can only be successful if people want to use it. Transparency about the selection criteria of the collection helps prospective readers and aggregators to determine whether the books on offer are of interest to them. In the case of OAPEN, the criteria (quality controlled monographs) are listed on the homepage, and the quality control process of the publishers is described.[2] The Directory of Open Access Books has adopted a similar policy.[3]

When a reader or an aggregator wants to use the content, the platform should make it easy to connect. We have seen before that the OAPEN Library is used via several channels: not just as an online public access catalogue, but also as a web based database that can be integrated into a larger collection. To ensure technical integration, the platform should offer its metadata based on standards that are used by the reader or aggregator. For instance, OAPEN supports aggregators with metadata feeds based on ONIX – a standard used in the publishing industry – and MARC21 – a library standard. Readers who are interested in a single title can download metadata in RIS format – to be used in citation managers – or use a widget to share the description via social media and mail. Connecting with readers or aggregators ought to go beyond technical measures. In the case of the OAPEN Library and DOAB, this is translated into agreements with commercial and non-commercial aggregators and by using social media to connect to individuals.

Language has proven to be an important influence on usage. Furthermore, the bulk of the usage of the OAPEN Library so far stems from the "global north". To extend the usage to the rest of the world, it might be useful to add monographs in Spanish and Portuguese to the collection – languages that are spoken in Latin-America. Also, a larger collection in French might be more attractive to the French-speaking countries in Africa.

The results so far are a good start towards understanding the effects of open access on monographs and the factors affecting usage of open access monographs. However, further research could help to deepen our understanding. The first research question would be the identification of usage by the general public. In my research, recognizing members of the public was based on eliminating possible organisational ties. Research on this topic should take into account privacy considerations; this has also been discussed in chapter 6.

2 http://oapen.org/content/peer-review-process-introduction

3 https://doabooks.org/doab?func=about&uiLanguage=en#purpose

In this publication, most of the research has been carried out on the OAPEN Library platform. As more open access book platforms are emerging, it will be interesting to repeat some of the experiments on those platforms. Wat are the effects of differences in technical abilities and book collections? Comparing usage results of multiple platforms has its own challenges; the COUNTER Code of Practice (COUNTER Online Metrics, 2014) might be useful in this case.

The effects of open access on the usage originating from developing countries has been discussed in detail. However, the collection of titles in this investigation have been provided by publishers from the "global north". If the collection of titles is enhanced with a sizable portion of titles from "global south" publishers, how would that affect the usage data? Does this lead to a higher percentage of usage from developing countries? Will the enhanced exposure be beneficial for authors?

I have deployed a clustering algorithm to find related books, based on usage by readers. The next phase would be to test several algorithms, in order to see if other procedures lead to comparable results. This will strengthen the claims of chapter 6. A related question is whether new algorithms lead to more fine-grained clusters.

Related to clustering algorithms, using text mining techniques to extract subjects from books might lead to new possibilities, for instance automatically clustering books based on distinctive words or word sequences and comparing these 'subject clusters' with the clusters of providers that were created for chapter 6.

Another possibility, based on the contents of the books, is to automatically define distinctive text segments, and searching whether they are used in newspapers, reports and other non-academic documents. This might help to determine the social impact of the monographs. The same technique could also be used as a service to readers, by searching for related academic open access documents in large databases such as BASE - Bielefeld Academic Search Engine.[4]

When the focus is widened beyond questions of usage, we might look a the role of paper books. Open access is inherently digital – based on online dissemination. Still, the role of paper books is not obsolete: the lack of influence of open access on sales of 'traditional' monographs points in that direction. Each publication form has its own merits, but it would be interesting to investigate whether the ideal of world wide free dissemination of knowledge can be combined with the affordances of paper publications.

4 https://www.base-search.net/about/en/

The success of this approach will depend on the stakeholders in scholarly publication.

Changes in online dissemination and the variations in stakeholder roles were already briefly discussed in section 10.4. The effects of this transition merit further research: if publishers continue to build online libraries, and academic libraries keep enlarging their publishing role, how will this affect scholarly communication?

In conclusion, the research on the dissemination of knowledge through open access monographs is far from finished. We have barely started.

11 References

About - Creative Commons. (n.d.). Retrieved from http://creativecommons.org/about

Abrizah, A., & Thelwall, M. (2014). Can the impact of non-Western academic books be measured? An investigation of Google Books and Google Scholar for Malaysia. *Journal of the Association for Information Science and Technology*, *65*(12), 2498–2508. https://doi.org/10.1002/asi.23145

Adema, J., & Ferwerda, E. (2014). Publication Practices in Motion : The Benefits of Open Access Publishing for the Humanities. In P. Dávidházi (Ed.), *New publication cultures in the humanities : Exploring the Paradigm Shift* (pp. 133–148). Amsterdam: Amsterdam University Press.

Ahmed, A. (2007). Open access towards bridging the digital divide–policies and strategies for developing countries. *Information Technology for Development*, *13*(4), 337–361. https://doi.org/10.1002/itdj.20067

Aleixandre-Benavent, R., Valderrama Zurián, J. C., Alonso-Arroyo, A., Miguel-Dasit, A., González de Dios, J., & de Granda Orive, J. (2007). [Spanish versus English as a language of publication and impact factor of Neurologia]. *Neurología (Barcelona, Spain)*, *22*(1), 19–26. Retrieved from http://europepmc.org/abstract/MED/17315099/reload=0

Alonso, C. J., Davidson, C. N., Unsworth, J. M., & Withey, L. (2003). *Crises and Opportunities: The Futures of Scholarly Publishing*. American Council of Learned Societies. Retrieved from http://www.acls.org/uploadedFiles/Publications/OP/57_Crises_and_Opportunites.pdf

American Library Association. (2014). Privacy : An Interpretation of the Library Bill of Rights. Retrieved March 5, 2017, from http://www.ala.org/advocacy/intfreedom/librarybill/interpretations/privacy

Antelman, K. (2004). Do Open-Access Articles Have a Greater Research Impact? *College & Research Libraries*, *65*(5), 372–382. https://doi.org/10.5860/crl.65.5.372

Archambault, É., Caruso, J., & Nicol, A. (2014). *State-of-art analysis of OA strategies to peer-review publications* (Vol. 1). Retrieved from http://science-metrix.com/files/science-metrix/publications/d_2.1_sm_ec_dg-rtd_oa_policies_in_the_era_update_v05p.pdf

Archambault, É., Côté, G., Struck, B., & Voorons, M. (2016). Research impact of paywalled versus open access papers. Retrieved from http://www.1science.com/oanumbr.html

Armstrong, C., & Ford, H. (2006). Africa and the digital information commons: An overview. *The Southern African Journal of Information and Communication*, *7*(7), 4–21. Retrieved from http://idl-bnc.idrc.ca/dspace/handle/10625/40002

Association of Research Libraries (ARL) :: ARL Statistics 2009-10. (2012). Retrieved May 14, 2012, from http://www.arl.org/stats/annualsurveys/arlstats/arlstats11.shtml

AUP. (2012). Amsterdam University Press. Retrieved November 17, 2011, from http://www.aup.nl

Bakos, J. Y. (1991). A strategic analysis of electronic marketplaces. *MIS Quarterly*, *15*(September), 295–310. https://doi.org/10.2307/249641

Bell, S., Shaw, B., & Boaz, A. (2011). Real-world approaches to assessing the impact of environmental research on policy. *Research Evaluation*, *20*(3), 227–237. https://doi.org/10.3152/095820211X13118583635792

Benneworth, P., & Jongbloed, B. W. (2009). Who matters to universities? A stakeholder perspective on humanities, arts and social sciences valorisation. *Higher Education*, *59*(5), 567–588. https://doi.org/10.1007/s10734-009-9265-2

Bennink, R., Meijer, I., Wamelink, F., & Zuijdam, F. (2008). *De maatschappelijke kwaliteit van onderzoek in kaart Een handreiking*. Utrecht. Retrieved from http://www.qanu.nl/comasy/uploadedfiles/MKO_handreiking_definitief.pdf

Berlin Declaration on Open Access to Knowledge in the Sciences and Humanities. (2003). In *Conference on Open Access to Knowledge in the Sciences and Humanities* (*20-22 Oct 2003, Berlin*). Berlin: Max Planck Gesellschaft. Retrieved from http://openaccess.mpg.de/286432/Berlin-Declaration

Bernal, I. (2013). Open Access and the Changing Landscape of Research Impact Indicators: New Roles for Repositories. *Publications*, *1*(2), 56–77. https://doi.org/10.3390/publications1020056

BioMed Central Ltd. (2014). BioMed Central | Using BioMed Central's open access full-text corpus for text mining research. Retrieved from http://www.biomedcentral.com/about/datamining

Boldrin, M., & Levine, D. K. (2002). The Case Against Intellectual Property. *American Economic Review*, *92*(2), 209–212. https://doi.org/10.1257/000282802320189267

Bollen, J., Van de Sompel, H., Hagberg, A., & Chute, R. (2009). A principal component analysis of 39 scientific impact measures. *PloS One*, *4*(6), e6022. https://doi.org/10.1371/journal.pone.0006022

Bollen, J., Van de Sompel, H., & Rodriguez, M. A. (2008). Towards usage-based impact metrics. In *Proceedings of the 8th ACM/IEEE-CS joint conference on Digital libraries - JCDL '08* (p. 231). New York, New York, USA: ACM Press. https://doi.org/10.1145/1378889.1378928

Bonn, M., & Furlough, M. (Eds.). (2015). *Getting the word out : academic libraries as scholarly publishers*. Chicago: American Library Association. Retrieved from www.ala.org/acrl/sites/ala.org.acrl/.../9780838986981_getting_OA.pdf

Book Industry Communication. (2010). BIC Standard Subject Categories – an Overview. Retrieved February 9, 2012, from http://www.bic.org.uk/7/BIC-Standard-Subject-Categories/

Borgman, L. C. (1999). What Are Digital Libraries?: Competing Visions. *Information Processing & Management*, *35*, 227–243. Retrieved from http://libezproxy.open.ac.uk/login?url=http://search.ebscohost.com/login.aspx?direct=true&db=eric&AN=EJ589995&site=eds-live&scope=site

Bornmann, L. (2014). Alternative metrics in scientometrics: A meta-analysis of research into three altmetrics. *ArXiv Preprint ArXiv:1407.8010*, *103*(3), 1123–1144. Digital Libraries; Physics and Society. https://doi.org/10.1007/s11192-015-1565-y

Budapest Open Access Initiative. (2012). Ten years on from the Budapest Open Access Initiative: setting the default to open. Retrieved April 29, 2014, from http://www.budapestopenaccessinitiative.org/boai-10-recommendations

Bulger, M. E., Meyer, E. T., De la Flor, G., Terras, M., Wyatt, S., Jirotka, M., … Madsen, C. M. (2011). Reinventing Research? Information Practices in the Humanities. *SSRN Electronic Journal*, (April), 83. https://doi.org/10.2139/ssrn.1859267

Cabezas-Clavijo, Á., Robinson-García, N., Torres-Salinas, D., Jiménez-Contreras, E., Mikulka, T., Gumpenberger, C., … Gorraiz, J. (2013). Most borrowed is most cited? Library loan statistics as a proxy for monograph selection in citation indexes. Digital Libraries. Retrieved from http://arxiv.org/abs/1305.1488

Calver, M. C., & Bradley, J. S. (2010). Patterns of citations of open access and non-open access conservation biology journal papers and book chapters. *Conservation Biology : The Journal of the Society for Conservation Biology*, *24*(3), 872–80. https://doi.org/10.1111/j.1523-1739.2010.01509.x

Carmody, T. (2011). This Is Why We'll Never Have Innovative E-Books | WIRED. Retrieved October 29, 2017, from https://www.wired.com/2011/08/this-is-why-well-never-have-innovative-e-books/

Carroll, M. W. (2006). Creative Commons and the New Intermediaries. *Michigan State Law Review*, *2006*(1), 45–65.

Central Intelligence Agency. (n.d.). The World Factbook – COUNTRY COMPARISON :: INTERNET HOSTS. Retrieved December 5, 2016, from https://www.cia.gov/library/publications/the-world-factbook/rankorder/2184rank.html

Chan, L., & Costa, S. (2005). Participation in the global knowledge commons. *New Library World*, *106*(3/4), 141–163. https://doi.org/10.1108/03074800510587354

Chan, L., Cuplinskas, D., Eisen, M., Friend, F., Genova, Y., Guédon, J.-C., … Velterop, J. (2002). Budapest Open Access Initiative. Retrieved from http://www.soros.org/openaccess/read.shtml

Chellappa, R. K., & Sin, R. G. (2005). Personalization versus privacy: An empirical examination of the online consumer's dilemma. *Information Technology and Management*, *6*(2–3), 181–202.

Chodorow, S. (1999). The Specialized Scholarly Monograph in Crisis: Or How Can I Get Tenure If You Won't Publish My Book? Retrieved March 12, 2012, from http://www.arl.org/resources/pubs/specscholmono/chodorow~print.shtml

Christian, G. E. (2008). Open Access Initiative and the Developing World. *African Journal of Library, Archives and Information Science*, *18*(2), 1–22. Retrieved from http://ssrn.com/paper=1304665

Collins, E., & Milloy, C. (2012). A snapshot of attitudes towards open access monograph publishing in the humanities and social sciences – part of the OAPEN-UK project. *Insights: The UKSG Journal*, *25*(2), 192–197. https://doi.org/10.1629/2048-7754.25.2.192

Collins, E., & Milloy, C. (2016). *OAPEN-UK final report: A five-year study into open access monograph publishing in the humanities and social sciences*. Retrieved from http://oapen-uk.jiscebooks.org/files/2016/01/OAPEN-UK-final-report-single-page-view.pdf

Corrado, E. M. (2007). Privacy and Library 2.0: How Do They Conflict? In *ailing into the future: charting our destiny: proceedings of the Thirteenth National Conference of the Association of College and Research Libraries*. Association of College and Research Libraries.

Costas, R., Zahedi, Z., & Wouters, P. (2014). *Do 'altmetrics' correlate with citations? Extensive comparison of altmetric indicators with citations from a multidisciplinary perspective* (CWTS Working Paper Series No. CWTS-WP-2014-001). Leiden. Retrieved from http://www.cwts.nl/pdf/CWTS-WP-2014-001.pdf

COUNTER Online Metrics. (2014). COUNTER | About Us. Retrieved March 1, 2015, from http://www.projectcounter.org/about.html

Cross, R. L. (2011). Digital books and the salvation of academic publishing. *The Bottom Line*, *24*(3), 162–166. https://doi.org/10.1108/08880451111185991

Daigle, L. (2004). WHOIS Protocol Specification. Retrieved March 22, 2015, from https://tools.ietf.org/html/rfc3912

Davis, F. D. (1989). Perceived Usefulness, Perceived Ease of Use, and User Acceptance of Information Technology. *MIS Quarterly*, *13*(3), 319. https://doi.org/10.2307/249008

Davis, F. D., Bagozzi, R. P., & Warshaw, P. R. (1989). User acceptance of computer technology: a comparison of two theoretical models. *Management Science*, *35*(8), 982–1003.

Diez, M. L. A., & Dempsey, L. (2006). The Library Catalogue in the New Discovery Environment: Some Thoughts. Retrieved March 7, 2017, from http://www.ariadne.ac.uk/issue48/dempsey

DOAB: Directory of Open Access Books. (n.d.). Retrieved November 29, 2012, from http://www.doabooks.org/

Emery, C., Lucraft, M., Morka, A., & Pyne, R. (2017). *The OA effect: How does open access affect the usage of scholarly books?* https://doi.org/10.6084/m9.figshare.5559280.v1

Ernø-Kjølhede, E., & Hansson, F. (2011). Measuring research performance during a changing relationship between science and society. *Research Evaluation*, *20*(2), 130–142. https://doi.org/10.3152/095820211X12941371876544

Falagas, M. E., Zarkali, A., Karageorgopoulos, D. E., Bardakas, V., & Mavros, M. N. (2013). The Impact of Article Length on the Number of Future Citations: A Bibliometric Analysis of General Medicine Journals. *PLoS ONE*, *8*(2). https://doi.org/10.1371/journal.pone.0049476

Ferwerda, E. (2014). Open access monograph business models. *Insights: The UKSG Journal*, *27*(s1), 35–38. https://doi.org/10.1629/2048-7754.46

Ferwerda, E., Snijder, R., & Adema, J. (2013). *OAPEN-NL - A project exploring Open Access monograph publishing in the Netherlands, Final Report*. The Hague. Retrieved from http://oapen.org/download?type=export&export=oapen-nl-final-report

Finch, J., Brindley, L., Blackman, T., Duffy, M., Waelde, C., England, J., … Hall, M. (2013). Open Access Publishing - Presentations from the Academy's 'Implementing Finch' Conference held on 29th and 30th November 2012 at the Royal Statistical Society in London. *Academy of Social Sciences Professional Briefings*, (01), 1–32.

Garfield, E. (2006). The history and meaning of the journal impact factor. *Jama*, *295*(1), 90–93.

Geyer-Schulz, A., Neumann, A., & Thede, A. (2003). Others also use: A robust recommender system for scientific libraries. In *International Conference on Theory and Practice of Digital Libraries* (pp. 113–125). Springer. Retrieved from http://www.em.uni-karlsruhe.de/

Ghosh, S. B., & Kumar Das, A. (2007). Open Access and Institutional Repositories A Developing Country Perspective: a case study of India. *IFLA Journal*, *33*(3), 229–250. https://doi.org/10.1177/0340035207083304

Gibbons, M. (1994). *The new production of knowledge : the dynamics of science and research in contemporary societies*. London [u.a.: Sage.

Giménez-Toledo, E., & Román-Román, A. (2009). Assessment of humanities and social sciences monographs through their publishers: a review and a study towards a model of evaluation. *Research Evaluation*, *18*(3), 13. https://doi.org/10.3152/095820209X471986

Glänzel, W., & Gorraiz, J. (2015). Usage metrics versus altmetrics: confusing terminology? *Scientometrics*, *102*(3), 2161–2164. https://doi.org/10.1007/s11192-014-1472-7

Glänzel, W., & Schoepflin, U. (1999). A bibliometric study of reference literature in the sciences and social sciences. *Information Processing & Management*, *35*(1), 31–44.

Gläser, J., Glänzel, W., & Scharnhorst, A. (2017). Towards a comparative approach to the identification of thematic structures in science. *Scientometrics*, *111*(2), 981–998. https://doi.org/10.1007/s11192-017-2295-0

Google Books. (n.d.). Reports for previews - Books Help. Retrieved April 18, 2012, from http://support.google.com/books/direct/bin/answer.py?hl=en-GB&answer=106172

Granados, N. F., Gupta, A., & Kauffman, R. J. (2006). The Impact of IT on Market Information and Transparency: A Unified Theoretical Framework. *Journal of the Association for Information Systems*, *7*(3), 148–178.

Grant, J., Brutscher, P.-B., Guthrie, S., Butler, L., & Wooding, S. (2010). *Capturing Research Impacts: A review of international practice*. Santa Monica: RAND. Retrieved from http://www.rand.org/pubs/documented_briefings/DB578.html

Greco, A. N., & Wharton, R. M. (2008). Should university presses adopt an open access [electronic publishing] business model for all of their scholarly books? In L. Chan & S. Mornati (Eds.), *ELPUB2008. Open Scholarship: Authority, Community, and Sustainability in the Age of Web 2.0 - Proceedings of the 12th International Conference on Electronic Publishing held in Toronto, Canada 25-27 June 2008* (pp. 149–164). Toronto. Retrieved from http://elpub.scix.net/data/works/att/149_elpub2008.content.pdf

Griffiths, J. R., & Brophy, P. (2005). Student searching behavior and the web: use of academic resources and Google. *Library Trends*, *53*(4), 539–554.

Guerrero-Bote, V. P., & Moya-Anegón, F. (2012). *Relationship between Usage and Citation and the influence of language*. Retrieved from http://ebrp.elsevier.com/pdf/2012_Proposal1-anegon_bote_morales.pdf

Guibault, L. (2011). Owning the Right to Open Up Access to Scientific Publications. In *Open Content Licensing : from Theory to Practice* (pp. 137–167). Amsterdam: Amsterdam University Press. Retrieved from http://www.oapen.org/record/389501

Hall, G. (2013). The Unbound Book: Academic Publishing in the Age of the Infinite Archive. *Journal of Visual Culture*, *12*(3), 490–507. https://doi.org/10.1177/1470412913502032

Hall, S. (2016). Will Google ever reopen signups for its Google Play Books self-publishing platform? Retrieved May 14, 2017, from https://9to5google.com/2016/12/28/will-google-ever-reopen-signups-for-its-google-play-books-self-publishing-platform/

Hammarfelt, B. (2014). Using altmetrics for assessing research impact in the humanities. *Scientometrics*, *101*(2), 1419–1430. https://doi.org/10.1007/s11192-014-1261-3

Harnad, S., Brody, T., Vallières, F., Carr, L., Hitchcock, S., Gingras, Y., … Hilf, E. R. (2004). The Access/Impact Problem and the Green and Gold Roads to Open Access. *Serials Review*, *30*(4), 310–314. https://doi.org/10.1016/j.serrev.2004.09.013

Harnad, S., Brody, T., Vallières, F., Carr, L., Hitchcock, S., Gingras, Y., … Hilf, E. R. (2008). The Access/Impact Problem and the Green and Gold Roads to Open Access: An Update. *Serials Review*, *34*(1), 36–40. https://doi.org/10.1016/j.serrev.2007.12.005

Harzing, A., & van der Wal, R. (2008). Google Scholar as a new source for citation analysis. *Ethics in Science and Environmental Politics*, *8*(1), 61–73. https://doi.org/10.3354/esep00076

Haustein, S., Larivière, V., Thelwall, M., Amyot, D., & Peters, I. (2014). Tweets vs. Mendeley readers: How do these two social media metrics differ? *IT-Information Technology*, *56*(5), 207–215.

Herb, U. (2010). Alternative Impact Measures for Open Access Documents? An examination how to generate interoperable usage information from distributed open access services. In *WORLD LIBRARY AND INFORMATION CONGRESS 76TH IFLA GENERAL CONFERENCE AND ASSEMBLY* (Vol. in press, p. 165–178 ST–OpenAccess Statistics: Alternative I). Retrieved from https://www.ifla.org/past-wlic/2010/72-herb-en.pdf

Herb, U., Kranz, E., Leidinger, T., & Mittelsdorf, B. (2010). How to assess the impact of an electronic document? And what does impact mean anyway?: Reliable usage statistics in heterogeneous repository communities. *OCLC Systems Services*, *26*(2), 133–145. https://doi.org/10.1108/10650751011048506

Hietanen, H. A. (2008). Creative Commons' Approach to Open Content. *SSRN Electronic Journal*, 1–88. https://doi.org/10.2139/ssrn.1162219

Hilton III, J. L., Lutz, N., & Wiley, D. (2012). Examining the reuse of open textbooks. *The International Review of Research in Open and Distance Learning*, *13*(2), 45–58.

Hilton III, J., Wiley, D., & Hilton, J. (2011). Free E-Books and Print Sales. *Journal of Electronic Publishing* (*JEP*), *14*(1). https://doi.org/10.3998/3336451.0014.109

Hilton, J. (2011). Free E-Books and Print Sales. *The Journal of Electronic Publishing*, *14*(1). https://doi.org/10.3998/3336451.0014.109

Holmberg, K., & Thelwall, M. (2014). Disciplinary differences in Twitter scholarly communication. *Scientometrics*, *101*(2), 1027–1042. https://doi.org/10.1007/s11192-014-1229-3

Hong, L., Convertino, G., & Chi, E. H. (2011). Language Matters In Twitter: A Large Scale Study. In *Fifth International AAAI Conference on Weblogs and Social Media Language* (pp. 518–521).

Houghton, J., Rasmussen, B., Sheehan, P., Oppenheim, C., Morris, A., Creaser, C., … Gourlay, A. (2009). Economic implications of alternative scholarly publishing models: Exploring the costs and benefits. *Victoria*, (January), 287. Retrieved from http://ie-repository.jisc.ac.uk/278/1/EI-ASPM_Summary.doc

Howard, B. (2012). Eric Hellman of Unglue.it on e-books, the creative commons, passionate authors and life after Amazon. Retrieved July 18, 2016, from http://teleread.com/eric-hellman-of-unglue-it-on-e-books-the-creative-commons-passionate-authors-and-life-after-amazon/

International Federation of Library Associations and Institutions. (2016). IFLA Code of Ethics for Librarians and other Information Workers (full version). Retrieved March 5, 2017, from http://www.ifla.org/publications/node/11092#privacy

International Monetary Fund. (2010). World Economic Outlook Database April 2010 -- WEO Groups and Aggregates Information. Retrieved from http://www.imf.org/external/pubs/ft/weo/2010/01/weodata/groups.htm

Jackson, R. (2014). The publisher journey for OUP. *Insights: The UKSG Journal*, *27*(s1), 21–25. https://doi.org/10.1629/2048-7754.117

Jaeger, P. T., McClure, C. R., Bertot, J. C., & Snead, J. T. (2004). The USA PATRIOT Act, the Foreign Intelligence Surveillance Act, and information policy research in libraries: Issues, impacts, and questions for libraries and researchers. *The Library Quarterly*, *74*(2), 99–121.

Jeckmans, A. J. P., Beye, M., Erkin, Z., Hartel, P., Lagendijk, R. L., & Tang, Q. (2013). Privacy in recommender systems. In *Social media retrieval* (pp. 263–281). Springer.

JISC - OAPEN-UK. (n.d.). Retrieved November 29, 2012, from http://oapen-uk.jiscebooks.org/

Jones, C. E. (n.d.). The ancient world online. Retrieved October 5, 2017, from http://ancientworldonline.blogspot.com/search?q=oapen

Jump, P. (2011). Monographs finally join citations database. *Times Higher Education*, (October 2011), 2013. Retrieved from http://www.timeshighereducation.co.uk/417771.article

Kelty, C. (2016). It's the Data, Stupid: What Elsevier's purchase of SSRN also means. Retrieved July 25, 2016, from http://savageminds.org/2016/05/18/its-the-data-stupid-what-elseviers-purchase-of-ssrn-also-means/

Kim, M. (2007). The Creative Commons and Copyright Protection in the Digital Era: Uses of Creative Commons Licenses. *Journal of Computer-Mediated Communication*, *13*(1), 187–209. https://doi.org/10.1111/j.1083-6101.2007.00392.x

Knöchelmann, M. (2017). Open Access Book Publishing and the Prisoner's Dilemma : A Theoretical Approach to a Description of the Slow Scalability of Open Access Book Publishing. In *BOOC*. UCL Press. https://doi.org/10.14324/111.9781911307679.11

Knowledge Exchange. (2010). The Impact of Open Access Outside European Universities. *Business*, 13. Retrieved from http://www.knowledge-exchange.info/Default.aspx?ID=412

Kortekaas, S., & Kramer, B. (2014). Thinking the unthinkable – doing away with the library catalogue. *Insights: The UKSG Journal*, *27*(3), 244–248. https://doi.org/10.1629/2048-7754.174

Kousha, K., & Thelwall, M. (2009). Google book search: Citation analysis for social science and the humanities. *Journal of the American Society for Information Science and Technology*, *60*(8), 1537–1549. https://doi.org/10.1002/asi.21085

Kousha, K., Thelwall, M., & Rezaie, S. (2011). Assessing the citation impact of books: The role of Google Books, Google Scholar, and Scopus. *Journal of the American Society for Information Science and Technology*, *62*(11), 2147–2164. https://doi.org/10.1002/asi.21608

Kumar, R., Raghavan, P., Rajagopalan, S., & Tomkins, A. (1999). Trawling the Web for emerging cyber-communities. *Computer Networks*, *31*(11–16), 1481–1493. https://doi.org/10.1016/S1389-1286(99)00040-7

Lamothe, A. A. R. A. (2010). Electronic Book Usage Patterns as Observed at an Academic Library: Searches and Viewings. *Partnership: The Canadian Journal of Library and Information Practice and Research*, *5*(1), 1–16. Retrieved from http://journal.lib.uoguelph.ca/index.php/perj/article/viewArticle/1071

Landes, W. M., & Posner, R. A. (1989). An Economic Analysis of Copyright Law. *Journal of Legal Studies*, *18*(2), 325.

Landry, R., Amara, N., & Lamari, M. (2001). Climbing the Ladder of Research Utilization: Evidence from Social Science Research. *Science Communication*, *22*(4), 396–422. https://doi.org/10.1177/1075547001022004003

Landry, R., Lamari, M., & Amara, N. (2003). The Extent and Determinants of the Utilization of University Research in Government Agencies. *Public Administration Review*, *63*(2), 192–205. https://doi.org/10.1111/1540-6210.00279

Lessig, L. (2004). The Creative Commons. *Montana Law Review*, *65*(1), 1–13.

Leydesdorff, L., & Etzkowitz, H. (1996). Emergence of a Triple Helix of University-Industry-Government Relations. *Science and Public Policy*, *23*, 279–286. Retrieved from http://dare.uva.nl/record/16759

Library Partners – Google Books. (n.d.). Retrieved April 3, 2016, from https://www.google.com/googlebooks/library/partners.html

Linden, G., Smith, B., & York, J. (2003). Amazon.com recommendations: Item-to-item collaborative filtering. *IEEE Internet Computing*, *7*(1), 76–80. https://doi.org/10.1109/MIC.2003.1167344

Linmans, A. J. M. (2009). Why with bibliometrics the Humanities does not need to be the weakest link. *Scientometrics*, *83*(2), 337–354. https://doi.org/10.1007/s11192-009-0088-9

Loren, L. P. (2007). Building a Reliable Semicommons of Creative Works: Enforcement of Creative Commons Licenses and Limited Abandonment of Copyright. *George Mason Law Review*, *14*(100), 271–328. https://doi.org/10.2139/ssrn.957939

Lyall, C., Bruce, A., Firn, J., Firn, M., & Tait, J. (2004). Assessing end-use relevance of public sector research organisations. *Research Policy*, *33*(1), 73–87. https://doi.org/10.1016/S0048-7333(03)00090-8

Lynch, C. (2002). Digital Collections, Digital Libraries and the Digitization of Cultural Heritage Information. *First Monday*, *7*(5), 1–13. https://doi.org/10.5210/fm.v18i5.4366

Maron, N., Mulhern, C., Rossman, D., & Schmelzinger, K. (2016). *The Costs of Publishing Monographs : Toward a Transparent Methodology*. Retrieved from http://www.sr.ithaka.org/publications/the-costs-of-publishing-monographs/

McCollough, A. (2017). Does It Make a Sound: Are Open Access Monographs Discoverable in Library Catalogs? *Portal: Libraries and the Academy*, *17*(1), 179–194. https://doi.org/10.1353/pla.2017.0010

Mendez, M., & Chapman, K. (2006). The use of scholarly monographs in the journal literature of Latin American history. *Electronic Journal of Academic and Special Librarianship*, *7*(3). Retrieved from http://southernlibrarianship.icaap.org/content/v07n03/mendez_m01.htm

Moghaddam, G. G., & Moballeghi, M. (2007). The importance of aggregators for libraries in the digital era. *Interlending & Document Supply*, *35*(4), 222–225. https://doi.org/10.1108/02641610710837536

Mönnich, M., & Spiering, M. (2008). Adding Value to the Library Catalog by Implementing a Recommendation System. *D-Lib Magazine*, *14*(5/6). https://doi.org/10.1045/may2008-monnich

Morisson, H. (2016). Dramatic Growth of Open Access September 30, 2016. Retrieved May 22, 2017, from http://poeticeconomics.blogspot.com/2016/10/dramatic-growth-of-open-access.html

Morrison, H. (2012). *Freedom for scholarship in the internet age*. Simon Fraser University. Retrieved from http://summit.sfu.ca/item/477

Mounier, P. (2011). Freemium as a sustainable economic model for open access electronic publishing in humanities and social sciences. *Information Services and Use*, *31*(3), 225–233. https://doi.org/10.3233/ISU-2012-0652

Nederhof, A. J. (2006). Bibliometric monitoring of research performance in the Social Sciences and the Humanities: A Review. *Scientometrics*, *66*(1), 81–100. https://doi.org/10.1007/s11192-006-0007-2

Nederhof, A. J. (2011). A bibliometric study of productivity and impact of modern language and literature research. *Research Evaluation*, *20*(2), 117–129. https://doi.org/10.3152/095820211X12941371876508

Neslin, S. A., Grewal, D., Leghorn, R., Shankar, V., Teerling, M. L., Thomas, J. S., & Verhoef, P. C. (2006). Challenges and Opportunities in Multichannel Customer Management. *Journal of Service Research*, *9*(2), 95–112. https://doi.org/10.1177/1094670506293559

Neslin, S. A., & Shankar, V. (2009). Key Issues in Multichannel Customer Management: Current Knowledge and Future Directions. *Journal of Interactive Marketing*, *23*(1), 70–81.

Newman, M. E. J., & Girvan, M. (2004). Finding and evaluating community structure in networks. *Physical Review E*, *69*(2), 026113. https://doi.org/10.1103/PhysRevE.69.026113

Nijboer, J. (2004). Big Brother versus anonymity on the Internet: implications for Internet service providers, libraries and individuals since 9/11. *New Library World*, *105*(7/8), 256–261. https://doi.org/10.1108/03074800410551002

Norris, M., Oppenheim, C., & Rowland, F. (2008). Open Access Citation Rates and Developing Countries. In *ELPUB 2008 Conference on Electronic Publishing* (pp. 335–342). Toronto, Canada. Retrieved from http://elpub.scix.net/data/works/att/335_elpub2008.content.pdf

OAPEN-UK. (2014). *Researcher survey 2014 » OAPEN-UK*. Retrieved from http://oapen-uk.jiscebooks.org/research-findings/researcher-survey-2014/

OAPEN.nl website - English. (n.d.). Retrieved November 29, 2012, from http://www.oapen.nl/index.php?option=com_content&view=article&id=58:english&catid=49:english&Itemid=63

OAPEN Consortium. (2011). *OAPEN Final Report*. Retrieved from http://project.oapen.org/images/documents/oapen_final_public_report.pdf

OAPEN Foundation. (2016). Organisation | OAPEN. Retrieved May 31, 2016, from http://oapen.org/content/organisation

Ochsner, M., Hug, S. E., & Daniel, H.-D. (2012). Indicators for Research Quality for Evaluation of Humanities Research : Opportunities and Limitations. *Bibliometrie - Praxis Und Forschung*, *1*(4), 1–17.

Ochsner, M., Hug, S., & Galleron, I. (2017). The future of research assessment in the humanities: bottom-up assessment procedures. *Palgrave Communications*, *3*, 17020. https://doi.org/10.1057/palcomms.2017.20

OCLC. (2016). OCLC WorldCat Discovery - Open access collections in WorldCat KnowledgeBase, 1–4. Retrieved from http://www.oclc.org/content/dam/oclc/worldcat-discovery/openaccess.pdf

Open Access Publishing in European Networks. (2010a). About OAPEN - Open Access Publishing in European Networks. Retrieved from http://project.oapen.org/about_OAPEN.asp

Open Access Publishing in European Networks. (2010b). OAPEN Library. Retrieved April 24, 2013, from http://www.oapen.org

Papin-Ramcharan, J., & Dawe, R. A. (2006). The Other Side of the Coin for Open Access Publishing – A Developing Country View. *Libri*, *56*(1), 16–27. https://doi.org/10.1515/LIBR.2006.16

Pazzani, M. M. J., & Billsus, D. (2007). Content-based recommendation systems. In *The adaptive web* (pp. 325–341). Springer.

Pinter, F. (2012). Open Access for Scholarly Books? *Publishing Research Quarterly*, *28*(3), 183–191. https://doi.org/10.1007/s12109-012-9285-0

Pinter, F. Google Scholar Indexes Open Access Books – Knowledge Unlatched (2015). Retrieved from http://www.knowledgeunlatched.org/2015/10/google-scholar-open-access-books/

Pinter, F. (2018). Why Book Processing Charges (BPCs) Vary So Much. *Journal of Electronic Publishing*, *21*(1). https://doi.org/http://dx.doi.org/10.3998/3336451.0021.101

Pochoda, P. (2013). The big one: The epistemic system break in scholarly monograph publishing. *New Media & Society*, *15*(3), 359–378. https://doi.org/10.1177/1461444812465143

Podlubny, I. (2005). Comparison of scientific impact expressed by the number of citations in different fields of science. *Scientometrics*, *64*(1), 95–99. https://doi.org/10.1007/s11192-005-0240-0

Poynder, R. (2014). Open and Shut?: The Open Access Interviews: Paul Royster, Coordinator of Scholarly Communications, University of Nebraska-Lincoln. Retrieved September 7, 2014, from http://poynder.blogspot.co.uk/2014/08/the-open-access-interviews-paul-royster.html

Priem, J., Piwowar, H. A., & Hemminger, B. M. (2012). Altmetrics in the wild: Using social media to explore scholarly impact. Digital Libraries. Retrieved from http://arxiv.org/abs/1203.4745

Priem, J., Taraborelli, D., Groth, P., & Neylon, C. (2011). altmetrics: a manifesto – altmetrics.org. Retrieved December 6, 2011, from http://altmetrics.org/manifesto/

Prins, A., Costas, R., Leeuwen, T. Van, & Wouters, P. (2014). Using Google Scholar in research evaluation of social science programs , with a comparison with Web of Science data. In *STI - Proceedings of the science and technology indicators conference 2014, Leiden "Context Counts: Pathways to Master Big and Little Data"* (pp. 434–443). Universiteit Leiden - CWTS. Retrieved from http://sti2014.cwts.nl/download/f-y2w2.pdf

Provan, K. G., Veazie, M. A., Staten, L. K., & Teufel-Shone, N. I. (2005). The use of network analysis to strengthen community partnerships. *Public Administration Review*, *65*(5), 603–613.

Redhead, C. (2012). Why CC-BY? - OASPA. Retrieved September 7, 2014, from http://oaspa.org/why-cc-by/

Ricci, F., Rokach, L., Shapira, B., & Kantor, P. B. (Eds.). (2011). *Recommender Systems Handbook*. Boston, MA: Springer US. https://doi.org/10.1007/978-0-387-85820-3

Rogers, E. M. (1995). *Diffusion of innovations*. New York [etc.]: The Free Press.

Royal Netherlands Academy of Arts and Sciences. (2010). *Quality indicators for research in the humanities*. *Humanities*. Amsterdam. Retrieved from https://www.knaw.nl/shared/resources/actueel/publicaties/pdf/quality-indicators-for-research-in-the-humanities

Salager-Meyer, F. (2008). Scientific publishing in developing countries: Challenges for the future. *Journal of English for Academic Purposes*, *7*(2), 121–132. https://doi.org/10.1016/j.jeap.2008.03.009

Scale, M.-S. (2008). Facebook as a social search engine and the implications for libraries in the twenty-first century. *Library Hi Tech*, *26*(4), 540–556. https://doi.org/10.1108/07378830810920888

Schafer, J. Ben, Konstan, J., & Riedl, J. (1999). Recommender systems in e-commerce. In *Proceedings of the 1st ACM conference on Electronic commerce* (pp. 158–166). ACM.

Schaffer, T. (2004). Psychology citations revisited: behavioral research in the age of electronic resources. *The Journal of Academic Librarianship*, *30*(5), 354–360. https://doi.org/10.1016/j.acalib.2004.06.009

Serenko, A., Bontis, N., & Moshonsky, M. (2011). Exploring the Role of Books as a Knowledge Translation Mechanism: Citation Analysis and Author Survey. In *Americas, The* (p. 11). Retrieved from http://aisel.aisnet.org/amcis2011_submissions/23/

Shen, Y. (2007). Information Seeking in Academic Research : *Information Technology and Libraries*, *26*(March), 4–14. Retrieved from http://search.ebscohost.com/login.aspx?direct=true&db=rzh&AN=2009564395&loginpage=Login.asp&site=ehost-live

Shneiderman, B., & Dunne, C. (2013). Interactive Network Exploration to Derive Insights: Filtering, Clustering, Grouping, and Simplification. In *Lecture Notes in Computer Science* (*including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics*) (Vol. 7704 LNCS, pp. 2–18). https://doi.org/10.1007/978-3-642-36763-2_2

Snijder, R. (2010). The profits of free books: An experiment to measure the impact of open access publishing. *Learned Publishing*, *23*(4), 293–301. https://doi.org/10.1087/20100403

Snijder, R. (2013a). A higher impact for open access monographs: disseminating through OAPEN and DOAB at AUP. *Insights: The UKSG Journal*, *26*(1), 55–59. https://doi.org/10.1629/2048-7754.26.1.55

Snijder, R. (2013b). Measuring monographs: A quantitative method to assess scientific impact and societal relevance. *First Monday*, *18*(5). https://doi.org/10.5210/fm.v18i5.4250

Snijder, R. (2014a). Modes of access: the influence of dissemination channels on the use of open access monographs. *Information Research*, *19*(3), 166–183. Retrieved from http://www.informationr.net/ir/19-3/paper638.html

Snijder, R. (2014b). The Influence of Open Access on Monograph Sales : The experience at Amsterdam University Press. *LOGOS: The Journal of the World Book Community*, *25*(3), 13–23. https://doi.org/10.1163/1878-4712-11112047

Snijder, R. (2015). Evaluating the Impact of the FWF-E-Book-Library Collection in the OAPEN Library: An Analysis of the 2014 Download Data. *D-Lib Magazine*, *21*(7/8). https://doi.org/10.1045/july2015-snijder

SNSF. (2015). Pilot project OAPEN-CH - SNF. Retrieved May 9, 2017, from http://www.snf.ch/en/funding/science-communication/oapen-ch/Pages/default.aspx

Spaapen, J., & van Drooge, L. (2011). Introducing "productive interactions" in social impact assessment. *Research Evaluation*, *20*(3), 211–218. https://doi.org/10.3152/095820211X12941371876742

SPARC Europe. (2015). The Open Access Citation Advantage Service. Retrieved October 29, 2015, from http://sparceurope.org/oaca

Springer. (2017). Book Production Workflow. Retrieved October 29, 2017, from http://www.springer.com/authors/book+authors/helpdesk?SGWID=0-1723113-12-803305-0

Steele, C. (2008). Scholarly Monograph Publishing in the 21st Century: The Future More Than Ever Should Be an Open Book. *The Journal of Electronic Publishing*, *11*(2). https://doi.org/10.3998/3336451.0011.201

Sterling, G. (2013). Topsy Becomes Definitive Twitter Search Engine. Retrieved from http://searchengineland.com/topsy-becomes-definitive-twitter-search-engine-171120

Stremersch, S., Verniers, I., & Verhoef, P. C. (2007). The Quest for Citations: Drivers of Article Impact. *Journal of Marketing*, *71*(3), 171–193. https://doi.org/10.1509/jmkg.71.3.171

Suber, P. (2008). Gratis and libre open access. *SPARC Open Access Newsletter*, (124). Retrieved from http://dash.harvard.edu/bitstream/handle/1/4322580/suber_oagratis.html?sequence=1

Suber, P. (2012). *Open Access*. Cambridge: MIT Press. https://doi.org/10.4271/2004-01-2697

Suber, P., Brown, P. O., Cabell, D., Chakravarti, A., Cohen, B., Delamothe, T., … Watson, L. (2003). Bethesda Statement on Open Access Publishing. *Access*, *3*(December 2012), 1–6. https://doi.org/10.4403/jlis.it-8628

Suzor, N. P. (2014). Free-riding, cooperation, and "peaceful revolutions" in copyright. *Harvard Journal of Law and Technology*, *28*(Fall), 0–74.

Swan, A., & Hall, M. (2010). Why Open Access can change science in the developing world. *Public Service Review: International Development Online*. Retrieved from http://usir.salford.ac.uk/id/eprint/9949

Tang, R. (2008). Citation Characteristics and Intellectual Acceptance of Scholarly Monographs. *College & Research Libraries*, *69*(4), 356–369. https://doi.org/10.5860/crl.69.4.356

Tennant, J. P., Waldner, F., Jacques, D. C., Masuzzo, P., Collister, L. B., & Hartgerink, C. H. J. (2016). The academic, economic and societal impacts of Open Access: an evidence-based review. *F1000Research*, *5*(632), 632. https://doi.org/10.12688/f1000research.8460.3

Thelwall, M. (2016). Interpreting correlations between citation counts and other indicators. *Scientometrics*, (Thelwall 2006). https://doi.org/10.1007/s11192-016-1973-7

Thelwall, M., Haustein, S., Larivière, V., & Sugimoto, C. R. (2013). Do altmetrics work? Twitter and ten other social web services. *PloS One*, *8*(5), e64841. https://doi.org/10.1371/journal.pone.0064841

Thelwall, M., & Sud, P. (2014). No citation advantage for monograph-based collaborations? *Journal of Informetrics*, *8*(1), 276–283. https://doi.org/10.1016/j.joi.2013.12.008

Thompson, J. B. (2005). *Books in the Digital Age: The Transformation of Academic and Higher Education Publishing in Britain and the United States*. Malden, Mass.: Polity Press.

Total number of Websites - Internet Live Stats. (n.d.). Retrieved July 11, 2018, from http://www.internetlivestats.com/total-number-of-websites/#trend

UNESCO. (2010). *UNESCO Science Report : The Current Status of Science around the World*. Paris: United Nations Education Scientific and Cultural Organisations. Retrieved from http://www.unesco.org/new/en/natural-sciences/science-technology/prospective-studies/unesco-science-report/unesco-science-report-2010/

Van Noorden, R. (2014). Elsevier opens its papers to text-mining. *Nature*, *506*(7486), 17–17. https://doi.org/10.1038/506017a

Vascellaro, J. E. (2009). Facebook, the Search Engine? - Digits - WSJ. Retrieved November 17, 2011, from http://blogs.wsj.com/digits/2009/08/11/facebook?-the-search-engine/

Venkatesh, V., & Davis, F. D. (2000). A theoretical extension of the technology acceptance model: four longitudinal field studies. *Management Science*, *46*(2), 186–204.

Verleysen, F. T., & Weeren, A. (2016). Mapping Diversity of Publication Patterns in the Social Sciences and Humanities: An Approach Making Use of Fuzzy Cluster Analysis. *Journal of Data and Information Science*, *1*(4), 33–59. https://doi.org/10.20309/jdis.201624

Wakita, K., & Tsurumi, T. (2007). Finding community structure in mega-scale social networks. *Proceedings of the 16th International Conference on World Wide Web*, 1275. https://doi.org/10.1145/1242572.1242805

Walker, S. R. (2009). Bioline International: A case study in open access and its usage for enhancement of research distribution for scientific research from developing countries. *OCLC Systems & Services*, *25*(2), 125–134. https://doi.org/10.1108/10650750910961929

Wang, P., Dervos, D. a., Zhang, Y., & Wu, L. (2007). Information-seeking behaviors of academic researchers in the Internet Age: A user study in the United States, China and Greece. In *Proceedings of the American Society for Information Science and Technology* (Vol. 44, pp. 1–29). https://doi.org/10.1002/meet.1450440273

Ware, M., & Mabe, M. (2015). *The STM report: An overview of scientific and scholarly journal publishing*. Retrieved from http://www.stm-assoc.org/2015_02_20_STM_Report_2015.pdf

Wasserman, M. (1998). Reprint: How Much Does It Cost to Publish A Monograph and Why? *The Journal of Electronic Publishing*, *4*(1). https://doi.org/10.3998/3336451.0004.104

Wasserman, S., & Faust, K. (1994). *Social network analysis: Methods and applications*. New York: Cambridge University Press. Retrieved from www.cambridge.org/9780521387071

White, H. D., Boell, S. K., Yu, H., Davis, M., Wilson, C. S., & Cole, F. T. H. (2009). Libcitations: A measure for comparative assessment of book publications in the humanities and social sciences. *Journal of the American Society for Information Science and Technology*, *60*(6), 1083–1096. https://doi.org/10.1002/asi.21045

WHOIS - Wikipedia. (n.d.). Retrieved November 23, 2011, from http://en.wikipedia.org/wiki/Whois

Williams, P., Stevenson, I., Nicholas, D., Watkinson, A., & Rowlands, I. (2009). The role and future of the monograph in arts and humanities research. *Aslib Proceedings: New Information Perspectives*, *61*(1), 67–82.

Willinsky, J., & Parry, F. (2006). The Access Principle: the case for open access to research and scholarship. *Access*, *49*, 165–168. Retrieved from http://www.equinoxjournals.com/LHS/article/viewPDFInterstitial/1629/2724

Winkmann, G., Schlutius, S., & Schweim, H. G. (2002). Citation Rates of Medical German-Language Journals in English-Language papers - Do They Correlate With the Impact Factor, and Who Cites? *DMW - Deutsche Medizinische Wochenschrift*, *127*(04), 138–143. https://doi.org/10.1055/s-2002-33307

Withey, L., Cohn, S., Faran, E., Jensen, M., Kiely, G., Underwood, W., … Keane, K. (2011). Sustaining Scholarly Publishing: New Business Models for University Presses. *Journal of Scholarly Publishing*, *42*(4), 397–441. https://doi.org/10.3138/jsp.42.4.397

World Bank. (2011). *The Little Data Book on Information and Communication Technology 2011. Communication* (Vol. 42). Washington: World Bank. https://doi.org/10.1596/978 - 0 - 8213 - 8248 - 6

World Bank, & Lewandowski, C. M. (2010). *The Little Data Book on Information and Communication Technology 2010. Communication* (Vol. 42). Washington: The World Bank. https://doi.org/10.1596/978-0-8213-8248-6

Wouters, P., Thelwall, M., Kousha, K., Waltman, L., De Rijcke, S., Rushforth, A., & Franssen, T. (2015). *The Metric Tide: Literature Review* (*Supplementary Report I to the Independent Review of the Role of Metrics in Research Assessment and Management*). https://doi.org/10.13140/RG.2.1.5066.3520

Zuccala, A. A., & White, H. D. (2015). Correlating Libcitations and Citations in the Humanities with WorldCat and Scopus Data. In *Proceedings of the 15th International Society for Scientometrics and Informetrics* (*ISSI*), *Istanbul, Turkey, 29th June to 4th July, 2015* (pp. 305–316). Istanbul: Bogazici University. Retrieved from http://forskningsbasen.deff.dk/Share.external?sp=S73e8693d-a836-422a-8d40-48eda088e60c&sp=Sku

Zuccala, A., Van Someren, M., & van Bellen, M. (2014). A machine-learning approach to coding book reviews as quality indicators: Toward a theory of megacitation. *Journal of the Association for Information Science and Technology*, *65*(11), 2248–2260. https://doi.org/10.1002/asi.23104

12 Appendix: published articles and data sets

This publication is article based. With the exception of chapter 6, the chapters have been published as journal articles in peer reviewed journals. The first article was published in 2013 in The Journal of Electronic Publishing; the most recent article was published in Scientometrics, in 2016. The texts have been incorporated unaltered, with one exception: all references have been combined in chapter 10.

Below is the list of published articles and the accompanying data sets

- Snijder, R. (2013). Do developing countries profit from free books?: Discovery and online usage in developed and developing countries compared. *The Journal of Electronic Publishing*, *16*(1), 1–14. https://doi.org/10.3998/3336451.0016.103.
- Snijder, R (2013). *3336451.0016.103-00000002* [Data set]. Retrieved from http://quod.lib.umich.edu/j/jep/images/3336451.0016.103-00000002.csv.
- Snijder, R. (2013). Measuring monographs: A quantitative method to assess scientific impact and societal relevance. *First Monday*, *18*(5). https://doi.org/10.5210/fm.v18i5.4250
- Ronald Snijder; OAPEN; (2012): *Measuring Monographs*. DANS. https://doi.org/10.17026/dans-24s-vcpz
- Snijder, R. (2014). The Influence of Open Access on Monograph Sales : The experience at Amsterdam University Press. *LOGOS: The Journal of the World Book Community*, *25*(3), 13–23. https://doi.org/10.1163/1878-4712-11112047.
- Snijder, R. (2014). Modes of access: the influence of dissemination channels on the use of open access monographs. *Information Research*, *19*(3), 166–183. Retrieved from http://www.informationr.net/ir/19-3/paper638.html
- Snijder, R. (2015). Better Sharing Through Licenses? Measuring the Influence of Creative Commons Licenses on the Usage of Open Access Monographs. *Journal of Librarianship and Scholarly Communication*, *3*(1), eP1187. https://doi.org/10.7710/2162-3309.1187
- Snijder, A.R. (OAPEN Foundation) (2013): *Better sharing through licenses*. DANS. https://doi.org/10.17026/dans-zpc-dmfb
- Snijder, R. (2016). Revisiting an open access monograph experiment: measuring citations and tweets 5 years later. *Scientometrics*, (May), 1–19. https://doi.org/10.1007/s11192-016-2160-6

- Snijder, MSc. R. (OAPEN Foundation) (2015): *Revisiting an Open Access monograph experiment: measuring citations and tweets five years later*. DANS. https://doi.org/10.17026/dans-x6m-67b2
- Snijder, R. (2017). *Patterns of information – clustering books and readers in open access libraries*. [Preprint] https://doi.org/10.17605/OSF.IO/UT23M
- Snijder, R. (OAPEN Foundation) (2017): *Patterns of information – clustering books and readers in open access libraries*. DANS. https://doi.org/10.17026/dans-x72-d9h2

ORCID ID:
0000-0001-9260-4941