Altmetrics to quantify the impact of scientific research published in open full text repositories

Bertil Dorch

Discussion paper

Altmetrics to quantify the impact of scientific research published in open full text repositories

S. B. F. Dorch
The Royal Library / University of Copenhagen Library, Copenhagen, Denmark
First version March 8, 2013

Abstract

Measures of the impact of scientific research increasingly become tools for various administrative applications. Bibliometric measures of the impact of research typically rely on citation analyses, which may not be appropriately representing all areas of scholarship, e.g. Arts and humanities. Metrics alternative to citation analyses are referred to as altmetrics, and include parameters derived from usage data, e.g. the number of times a paper is downloaded. In this paper I consider usage data from the open full text repository hprints.org for Nordic arts and humanities, and on the basis of an analysis of a small dataset from hprints.org, suggest new altmetrics based on the normalized squared number of downloads d downloads that could prove to be a valuable tool to quantify the impact of research published in open full text repositories.

1. Introduction

As quantitative measures of impact increasingly become common tools for research politics, funding and research evaluation, it is also becoming evident that disciplinary traditions regarding publication and reference patterns differ in ways that can and should lead to stakeholders questioning the use of impact measures derived solely from bibliometric measures. Such data may e.g. either not be available in quantities yielding statistical significance, or data may be incomparable among subjects, even if such data is both sound and plentiful.

Bibliometric measures of the impact of research typically rely on citation analyses of research publications, i.e. directly on citation data, e.g. Hirsch's h-index (Hirsch 2005) in the case of an individual's impact, or Garfield's Journal Impact Factor in the case of a publication channel's impact (e.g. Garfield 1998). Citation data used to perform citation analyses almost always stem from huge databases such as Thomson Reuters' Web of Knowledge or Elsevier's Scopus.

However, the fraction of arts and humanities currently available in Web of Knowledge - through the Arts & Humanities Citation Index (AHCI), is much lower than that of science, technology

and medicine, and Faurbæk (2007) concludes, that "the coverage of Danish articles in AHCI is only approx. 13 percent, which is considered to be an extremely low coverage."

According to Auken (2013), the STM area are normatively stronger than arts and humanities, when it comes to describing what is "true" scholarship, and "in order for bibliometrics to be even marginally acceptable within the humanities they need to be able to treat the humanities in a way that reflect the differences between the two fields." One of the said differences lies in genres, quoting again Auken (2013) "that even if you perceive scholarly research in the humanities as something that takes place in articles and monographs you are being reductive as for instance archival, museal or philological work is crucial to humanistic scholarship as well as are all the other publication forms." Other differences lie in the function of public outreach, the use of national languages as opposed to English, authoring and longevity.

Hence, it would seem that academic impact cannot generally be measured on the basis of citation data.

Generally, Research Councils UK[1] describes academic impact in the following ways: "The demonstrable contribution that excellent research makes to academic advances, across and within disciplines, including significant advances in understanding, methods, theory and application".

Demonstrability traditionally infers quantitative method, as in the case of bibliometrics or citation analysis. Metrics alternative to citation analyses and traditional bibliometrics are often referred to as altmetrics, and include e.g. metrics of use such as views and references in social media. An example of an implementation of altmetrics is *Public Library of Science* (PLoS[2]), who starting 2009 introduced article level metrics on all PLoS' journals, complying with a certain standard (COUNTER-3).

In this paper I suggest supplementing citation measures - particularly in the case of publications freely available in open archives - by altmetrics derived from usage statistics. Such altmetrics could prove particularly relevant to arts and humanities.

However, online usage data should be interpreted with caution. In general, it is dependent on the age of the article and its subject area - much like citations are. And on the one hand, "a read" does not always lead to "a quote". On the other hand, papers may have an impact, without being quoted or cited: it can impact the reader's mind, it may influence and help create new ideas, it may shape the way of thinking, within that particular field of research. Hence "reads" can in a way be considered to be much more fundamentally scholar academic impacts than citations.

[1] http://www.rcuk.ac.uk/kei/impacts/Pages/meanbyimpact.aspx
[2] "Article-Level Metrics Information".

2. Dataset

The dataset considered in this paper, was provided by hprints.org, which is an archive for electronic preprints of scientific papers in the fields of arts and humanities. Hence, the case is an archive that can be accessed freely via the Internet, i.e. an Open Access repository making research papers publicly available a potentially wide range of audiences (cf. Dorch *et al.* 2010).

The aim of hprints is to make Nordic research available through an Open Access online electronic full text archive, but the limitation to Nordic countries is claimed to be mainly an initial restriction for funding reasons. The archive will primarily contain electronic research documents in the form of preprints, reprints, working papers, book chapters, conference reports, invited lecture manuscripts etc. The archive is set up, maintained and promoted by Copenhagen University Library Service and consortium members from Nordic countries.

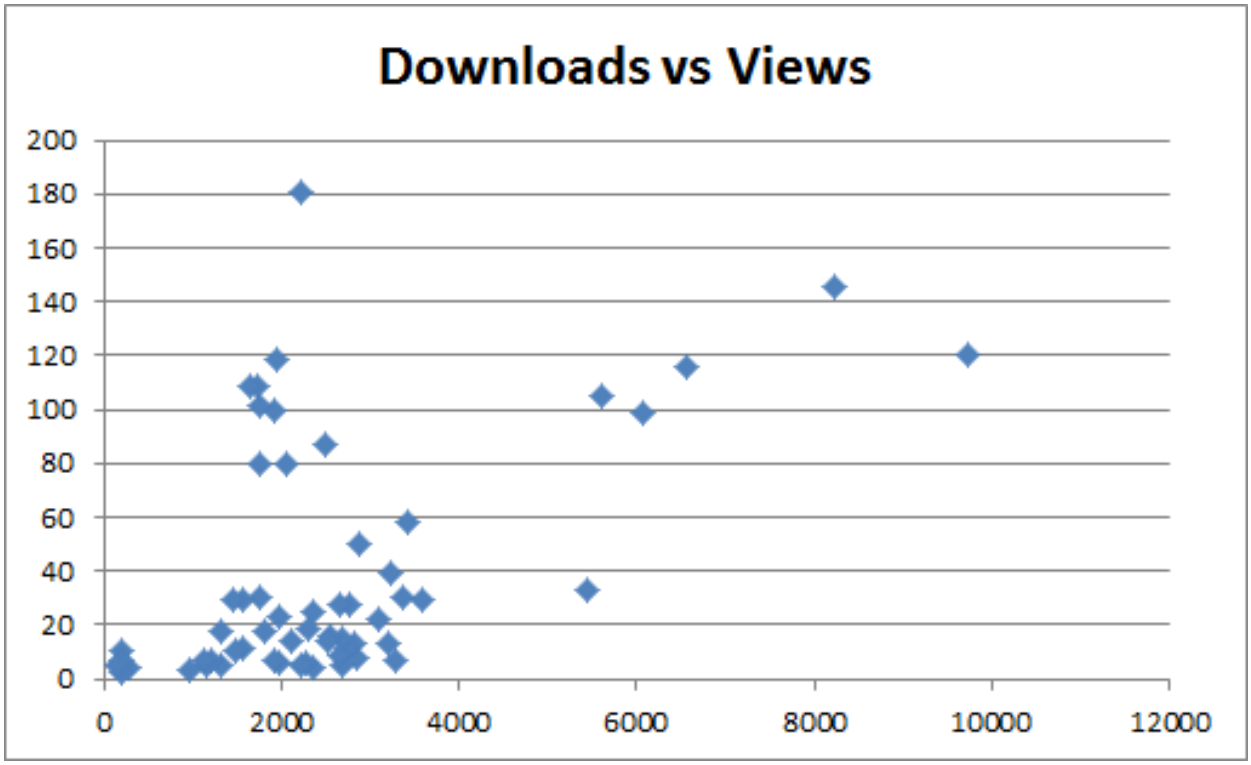

Fig. 1. A scatter plot of the dataset: Blue diamonds are the number full text downloads *Nd*, as a function of the number of page views, *Nv*.

hprints.org can be used as a tool for scientific communication between academic scholars, who can upload full-text research material such as articles, papers, conference papers, book chapters etc. The content of the deposited material should be comparable to that of a scientific paper that a scholar would consider suitable for publication in for example a peer reviewed scientific journal. It is possible to search and find the paper by defined topics through an Internet search. Secondly, all submitted papers are stored permanently and receive a stable web address.

The main data parameters provided are the number of page views (N_v) and the number full text downloads (N_d), for a chosen subset of 60 papers, deposited from January 2008 (the startup of hprints.org) until July 2012, cf. Fig. 1. The papers were chosen from a criterion that they had been downloaded more than once, that they were deposited more than half a year prior to the date of accessing the data (January 31th 2013), and that they did not seem otherwise biased due to e.g. extraordinary singular exposure on the Internet by third parties, say e.g. as known news features or similar. The data selection procedure involved reviewing the available statistics for each individual paper according to the above criteria, and no other selection criterion was applied.

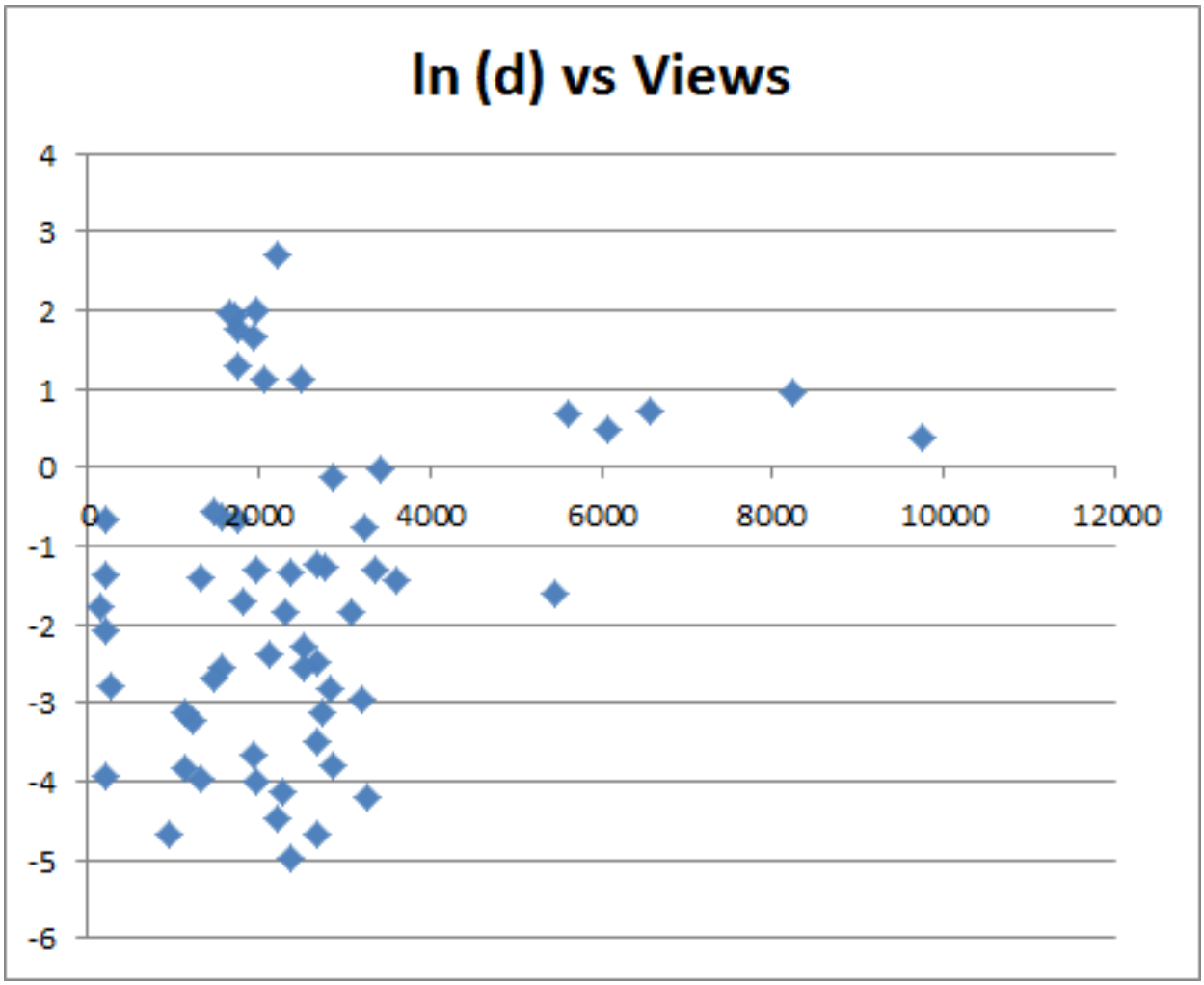

Fig. 2. Scatter plot of dataset showing the distribution of $ln\ (d)$ versus number of views, N_v.

3. Analysis and results

The derived usage metrics to be considered here, besides the basic usage parameters N_v and N_d, are the mean number of downloads per hit, $D = N_d / N_v$, and the normalized squared number of downloads, $d = D * N_d$.

Hence, $D = 1$ means one download pr. hit, which is an ideal and normally unrealistic situation only reserved for the absolute highest impact papers, where one view of the title, author, abstract or keywords at the landing page leads to a full text download. On the other hand, $D << 1$ means that views rarely leads to download.

When it comes to the normalized squared number of downloads $d = N_d$ corresponds to $D = 1$. However, since both the basic usage parameters N_v and N_d, generally are large - otherwise there would be no impact to infer, it is relevant to consider the natural logarithm of these numbers, or in terms of d:

$$ln\,(d) = 2*ln\,(N_d) - ln\,(N_v) = 2*ln\,(N_d) - 0.43*log\,(N_v). \quad (1)$$

Figure 2 shows $ln\,(d)$ as a function of the number of views N_v for the dataset: it appears that several papers have negative $ln\,(d)$ and relatively lower number of views, while some have positive $ln\,(d)$ and number of views spanning the full scale.

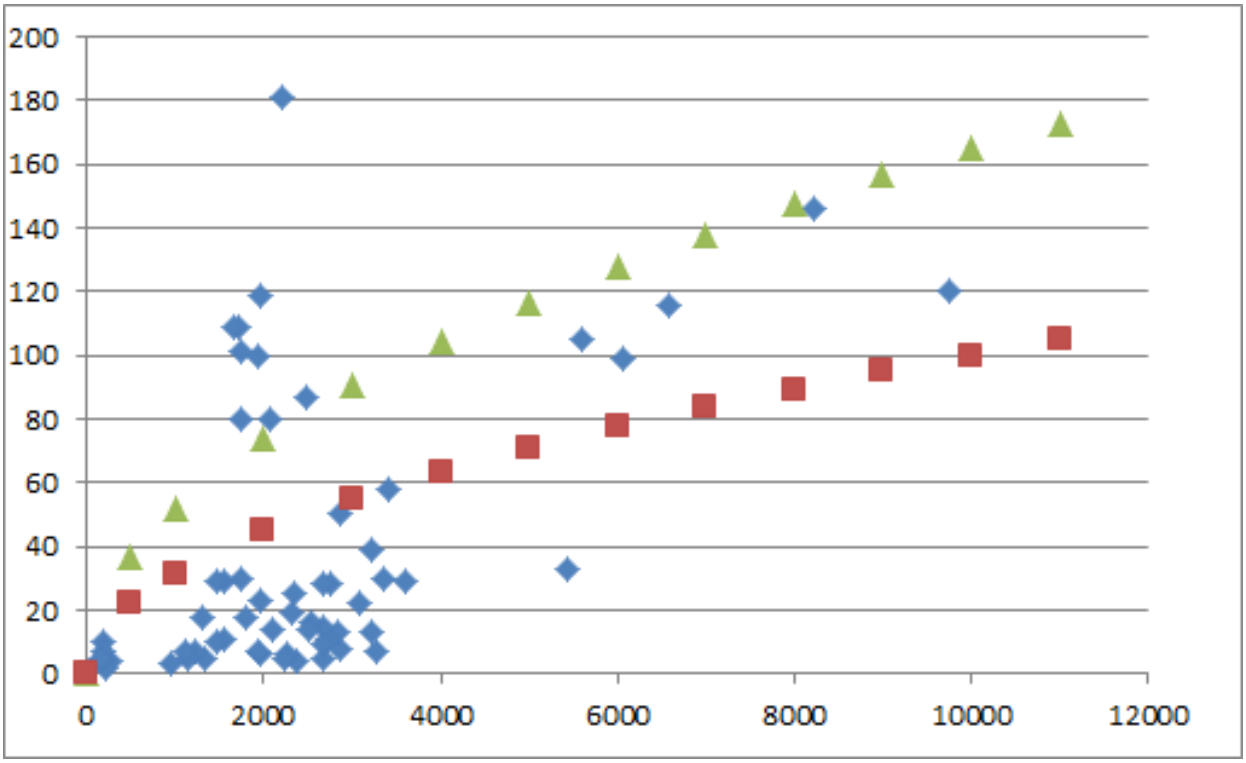

Fig. 3. Scatter plot of the dataset identical to Fig. 1: Blue diamonds are the number full text downloads (N_d) as a function of the number of page views (N_v). Red squares correspond to the lower limit of $ln\,(d) = 0$ in Eq. (2), and the green triangles to the upper limit of $ln\,(d) = 0$ in Eq. (3).

Indeed, a lower limiting case of $ln\,(d)$ is when the number of views N_v is quadratically larger than the number of downloads N_d, i.e. when $d = 1$, or when

$$ln\,(d) = 0 \Rightarrow 2*ln\,(N_d) = ln\,(N_v) \Rightarrow N_d = \text{sqrt}\,(N_v). \quad (2)$$

If the number of downloads per view D is high, the normalized squared number of downloads d can be higher than unity, and another limiting case is when

$$ln\,(d) = 1 => d = e => N_d\text{^}2 \,/\, N_v = e => N_d = \text{sqrt}\,(e^* N_d). \quad (3)$$

The idea now is that these two limits can be used to differentiate the "impact" of the papers in the dataset, i.e. as alternative metrics, cf. Fig. 3. I.e. for an individual paper, one may ask "how high is d?" or "Is d above 1, 2 or 3?"

Hence, d or $ln(d)$ could be useful altmetrical parameters when evaluating the impact of papers, since papers with $d > 1$ seem to imply that these papers belong to a family of papers that have higher download success than other papers. Similarly goes for papers with $ln\,(d) > 0$, though this may be a less convenient parameter to calculate and interpret in practice.

The above concerns mostly the measure of a paper's impact. When it comes to measuring an individual's impact, the usage parameter that ideally could be considered to equal the potential number of citations, is the average number of downloads per paper $< N_d >$, which turns out to be a useful parameter.

Table 1 below lists various usage parameters for seven papers by the same individual. On the one hand, it turns out that in this particular case at least, there are seemingly correlations per paper between d, or $\ln(d)$, and the number of downloads N_d per year, and the number of downloads per view D. On the other hand, there seems to be no similar trend between these numbers, and the number of views or downloads.

No.	deposit date	N_v	N_d	D^*100	d	$\ln(d)$	N_v / year	N_d / year
7	*2012-07*	2494	87	3,49	3,03	1,11	4275,4	149,1
3	*2008-03*	6572	116	1,76	2,05	0,72	1335,7	23,6
2	*2008-01*	6062	99	1,63	1,62	0,48	1192,5	19,5
5	*2010-01*	2866	50	1,74	0,87	-0,14	929,5	16,2
1	*2008-01*	3590	29	0,81	0,23	-1,45	706,2	5,7
4	*2008-05*	5440	33	0,61	0,20	-1,61	1145,3	6,9
6	*2011-11*	2226	5	0,22	0,01	-4,49	1780,8	4,0

Table. 1 Nine usage parameters for seven papers by the same individual ordered according rising d.

The table lists the seven papers by the same author according to rising d. In this particular case, one could argue that the average number of downloads per year could also be an interesting measure, however, it is less obvious what the impact threshold should be, and in any case there is

the question of news value, which may not be relevant to research with more long term relevance.

4. Results and discussion

We propose that analyses of usage statistics combining the number of page views N_v, with the number full text downloads N_d could be used as metrics to discern the impact of scientific research published in open full text repositories.

However, page views of landing pages - or "hits" - do not in itself imply impact. The same can be said to be true in the case of downloads, even though the probability that the particular research publication does in fact have an impact ceteris paribus is higher, when it is downloaded by an individual, since downloading would precede the process of reading, evaluating and applying the research. I.e. the landing page or placeholder of a full text paper may be viewed any number of times, leading to no impact, since the reader does not proceed to download it (and read it), while a "hit" that leads to a download could be considered to be a prerequisite for impact.

Perneger (2004) investigated research papers in *British Medical Journal* (BMJ) to examine a potential relation between "hits" and subsequent citations: for papers published in the BMJ in 1999, the hit count on the website in the week after online publication predicted the number of citations in subsequent years. Perneger (2004) concluded that "the hit count is a potentially useful measure of the scientific value of a research paper".

On the basis of the simple results from hprints.org presented here, it is tempting to argue that d or $ln\ (d)$ could be similarly useful altmetrical parameters when evaluating the impact of papers deposited in open archives, e.g. by asking for the number - or fraction - of an individual's publications with d larger than unity - or similarly, with $ln\ (d)$ larger than zero.

Naturally, the validity and usefulness of d as an altmetrical impact parameter should be evaluated by analyses carried out with a different and preferably larger dataset. In particular a future study should examine the potential correlation between d, D or $ln\ (d)$ and classical bibliometrical parameters, e.g. h-index and citation numbers.

However, as with self- of group citations, usage statistics suffers the potential bias resulting from authors playing the F5-reload game. Hence, it is very important for the future use of altmetrics that open archives systems enable repository managers - and third party tools such as Altmetrics.com etc. - to avoid misusage or in filtering usage data appropriately rendering altmetrics a trustworthy supplement to traditional citation bibliometrics.

Acknowledgements

This research has made use of hprints.org which is funded by Nordbib and hosted by Centre National de la Recherche Scientifique (CNRS). The author also wishes to thank the following colleagues for providing useful discussions improving this manuscript: Simone Schipp von Branitz Nielsen and Asger Væring Larsen.

References

Auken, S. (2013): "Measuring the Spirit? Bibliometrics and the Humanities". *ACUMEN Open Seminar The 8th of March 2013, Royal School of Library and Information Science, Copenhagen, Denmark*. http://hprints.org/hprints-00801367

Dorch, S.B.F. Rabow, I., Sikström, M., Schipp von Branitz Nielsen, S., Gilbert, J., Nyman, M. and Drachen, T.M. (2010): *ScieCom Info* **6** (2), p.1. http://hprints.org/hprints-00494135

Faurbæk, L. (2007): "Humanistisk Forskningskvalitet". Report. University of Copenhagen. http://hprints.org/hprints-00208083

Garfield, E. (1998): "The Impact Factor and Using It Correctly". *Der Unfallchirurg* **101** (6): p.413. PMID 9677838.

Hirsch, J.E. (2005): "An index to quantify an individual's scientific research output", *Proceedings of the National Academy of Science* **46**, p.16569: http://arxiv.org/abs/physics/0508025

Perneger, T. V (2004): "Relation between online "hit counts" and subsequent citations: Prospective study of research papers in the BMJ". *BMJ* **329** (7465), p.546.doi:10.1136/bmj.329.7465.546. PMC 516105. PMID 15345629.