Date submitted: 03/07/2010

Download data versus traditional impact metrics: Measuring impact in a sample of biomedical doctoral dissertations

Urban Andersson
Jonas Gilbert
and
Karin Henning
Gothenburg University Library
Gothenburg, Sweden

Meeting: **155. Science and Technology Libraries**

WORLD LIBRARY AND INFORMATION CONGRESS: 76TH IFLA GENERAL CONFERENCE AND ASSEMBLY
10-15 August 2010, Gothenburg, Sweden
http://www.ifla.org/en/ifla76

Abstract:

Download data and traditional impact metrics such as citation analysis both measure usage and possible impact of research. The case chosen for this analysis is based on dissertations of a cohort at the Sahlgrenska Academy, the faculty of Health Sciences of University of Gothenburg. At the Academy, a committee responsible for the PhD education selects the "Best doctoral thesis" of the year. This made it possible to carry out additional analysis, as examining the relationship between peer judgement and usage data (citations and downloads), and studying how promotional activities influence the usage of the dissertations. This praxis-oriented research investigates how a university library can develop different methods to provide valid user data. Analyses were carried out using field normalized citation scores and aggregated usage log data. Bivariate analysis measured the correlation between the data. The results showed no correlation between downloads and citations. However, there was a positive correlation between the peer judgments and field normalized citation scores, and between PR activities and usage through downloading.

Introduction

As librarians at Gothenburg University Library, working with bibliometrics and the institutional repository, we are experiencing growing interest from the researchers and faculties concerning usage data for e-published material. The underlying question that we are addressing is if and how this locally aggregated data can complement citation data. Our ambition has been to do a praxis-oriented study of this relationship with regard to the doctoral theses published by the Sahlgrenska Academy (the faculty of Health Sciences), which annually publishes approximately 150 PhD theses.

To give an outline of the study, we will start by describing the bibliometric service that the library operates and the routines for the e-publishing of doctoral theses at the university. After that we describe the case for the study in detail. We then look at some of the questions posed by recent research within this field, before accounting for the method and outcome of this study.

Bibliometric Services

Gothenburg University has approximately 37 000 students and 5 000 employees and covers most of the scientific disciplines (except technology and engineering, which resides at Chalmers University of Technology). Gothenburg University Library is part of the university and consists of seven libraries and three learning centers, serving the different faculties within the university. Among the central departments serving the library are Bibliometric Services and Digital Services. Bibliometric Services has been operating since 2008, staffed with three persons (FTE 1, 7), and a network of librarians with a special interest in bibliometrics. On the whole, bibliometrics is showing a growing interest in Sweden - partly as a result of the national bibliometric indicator for allocating research funds (Carlsson, 2009). During 2009 the department delivered about 50 analyses for different clients within the university. These bibliometric analyses cover a wide spectrum, including bibliometric indicators for local allocation, science mapping, collaboration and network analyses, different citation and performance analyses, and analyses for the library's collection development. The department hosts its own relational database, containing data from the local publication database GUP, citation data from ISI Web of Science (including field normalized scores) and Scopus, data from the Norwegian and Danish national systems for valuing publication channels, Google Scholar citations, data from the Swedish national library catalogue Libris and also data from Swepub, the national publication database. The database is designed by Håkan Carlsson (bibliometrician at Bibliometric services) and based on the database management system PostgreSQL.

Digital Services

The university library is responsible for and operates a database where all researchers at the university are mandated to register bibliographic data for the published research output.[1] The library also operates an e-publishing/open repository service, *Gothenburg University Publications Electronic Archive (GUPEA)* [2]. A team within the department Digital Services is responsible for the administration and support of the publication and e-publishing databases. The e-publishing service - GUPEA - uses the DSpace platform and was launched in 2006. DSpace is a software used for building open digital repositories. Developed originally by MIT and Hewlett-Packard, it is today one of most well-know and widespread open source initiatives within academia.[3] Gothenburg University Library is active within the DSpace user community and hosted an international user group meeting in 2009. As an indication of the overall impact of the e-publishing service, it could be noted that GUPEA qualifies as #35 on the Ranking Web of World Repositories.[4]

[1] *Gothenburg University Publications* (GUP), http://gup.ub.gu.se/gup/

[2] http://gupea.ub.gu.se/

[3] http://www.dspace.org

[4] Ranking Web of World Repositories: Top 400 Institutional Repositories., http://repositories.webometrics.info/top400_rep_inst.asp (June 2010)

E-publishing of doctoral dissertations

The content of the repository includes both textual output (theses, articles, books etc), as well as video- and audio-material. However, a main focus has been to establish routines within the university to collect and e-publish digital versions of the doctoral dissertations. In 2006, in close cooperation with the Sahlgrenska Academy, we established a workflow for the e-publishing of doctoral theses. These routine has since then been adopted by the other faculties of the university. The e-thesis workflow requires the PhD student to register bibliographic details before the public defence, and to upload files (pdf) containing the fulltext of the thesis. Within the university, both monographic theses and compilation theses are written. Compilation theses are used within biomedical sciences, science and some social science and humanities departments and are comprised of three to six articles, and a comprehensive summary. Monographic theses are written within the arts and humanities and a few other departments. In cases where there are copyright issues preventing e-publishing of the fulltext, for instance when a thesis also is published as a book by an external publisher, the author uploads an abstract. For compilation theses, it is the comprehensive summaries, in Swedish referred to as the "thesis frame", that are e-published in the repository. While the printed theses also include papers that are either already published in journals or in manuscript form (generally in a submitted or accepted status), the e-published version consists of the thesis frame in combination with bibliographic data and URLs (using either DOI or PubMed-id) for the published articles. Before the thesis becomes e-published and available on the Internet, the registration is approved by the administrative office of the faculty.[5] The e-published theses are linked to lists of forthcoming events. For a subset of the theses, the information officers write press releases in English. The press releases are published in international research news portals such as AlphaGalileo and IDW-online.[6]

Objectives

The objective of our research is to investigate the relationship between download data and traditional impact metrics such as citation analysis, above all in order to find out if the different methods are correlating or complementary. In this study, we also could compare (on a small basis) peer judgement with usage data (citations and downloads). Another investigated relationship is how marketing influences the usage of the dissertations. This praxis-oriented research examines how a university library can develop different methods to provide valid user data.

The case that we have chosen for the analysis is based on an event at the Sahlgrenska Academy where the committee responsible for the PhD education selects the "Best doctoral thesis" of the year. One thesis is selected as the most outstanding overall, and the six departments within the Academy select one thesis respectively. In other words, seven theses in all from the 158 published during the year are selected. In the first phase, the peers within the faculty nominate candidates for this election. The committee then collects various data for the nominated theses, and among these data are download data from the repository. For 2009, 22 theses were nominated. The election took place in May 2010 and we delivered download data from January 2009 to April 2010. The method for collecting this data is described below. However, it is obvious that a thesis published early in the year has had more time to collect usage data and this could be considered a methodological issue.

[5] For a detailed overview of the registration process, see "Instructions for E-publishing a PhD Thesis in GUPEA": http://www.ub.gu.se/publicera/epublicering/doktor/Doktorandanvisningar_eng_espikningMars09.pdf

[6] http://www.alphagalileo.org, http://idw-online.de/

Previous research

The context for our study is the deployment of usage log data as a complement to citation based metrics. As more and more of the scientific communication takes place in an online environment, the interest in using different forms of usage or download based metrics for evaluating research increases. Several authors have noted that the primary reason for this is the wish to overcome the time-delay associated with citation based metrics(Armbruster, 2008; Neylon & Wu, 2009). Armbruster also sees a connection to a wider purpose of evaluation: "Rather than asking how metrics may be used to evaluate scholars, the question should be: what kind of metric information services would serve scholars?" (Armbruster, 2010, p. 34).

Among the research projects addressing the relationship between citation metrics and usage log data, Metrics from Scholarly Usage of Resources (MESUR) is perhaps the most comprehensive. Based on a large collection of usage data from publishers, universities and other aggregators, the MESUR project has undertaken analyses to compare usage log data with citation based impact measures (Bollen, Van de Sompel, Hagberg, & Chute, 2009). The results from these analyses reinforce the need to use multiple measures to describe scientific impact. As for the service providers - either journal publishers or repository administrators - there are several initiatives to display usage data on a single item level. A recent example, launched in 2009, is the Public Library of Science article-level metrics program.[7] In connection to this interest in usage log data, the increasing need to standardise and to make usage data comparable and exchangeable has been emphasized (Merk, Scholze, & Windisch, 2009). The Coats (2005) study is based on articles in *International Journal of Cardiology*, where two “top-ten-lists”, one with the most downloaded and one with the most cited, were compared. The result showed no correlation between the different methods. This small sample can probably not be used to hint at wider implications, but what is interesting is the detection of different articles characteristics between the lists. The citation list was dominated by original articles, but the download list included “up-to-date reviews of either cutting-edge topics (such as the potential of stem cells) or the management of rare and unusual conditions”(Coats, 2005, p. 124). This study applies the same observation window for the citation data and downloads, which can be a problem. Among others, Schloegl and Gorraiz (2010) have presented different obsolescence patterns between usage and citation data. In their study of oncology journals it was indicated that the mean half-life of usage was 1,7 years, compared with 5,6 years for citations. The study was based on oncology journals from 2001 to 2006, and showed a moderate correlation between citations and downloads. From another perspective, Duy and Vaughan (2006) study how to use different methods for library collection management of journals. Their study indicates a positive correlation between the university staff’s usage of journals by downloading, and the staff’s usage by citing. However there was no relationship between the (local) usage data and (global) citation data as Journal Citation Reports (provided by ISI Web of Science).

Methodology and data collection

The download data was collected from the University's open repository, GUPEA, running on the DSpace platform. Usage events, such as page views, downloads and searches, are recorded in real-time and stored in a separate database. This data is periodically processed, in order to remove events caused by known robots, such as google-bot, based on recorded IP-

[7] http://www.plos.org/cms/node/485, http://article-level-metrics.plos.org/

numbers. We also attempt to remove events by other clients that behave in a bot-like manner, but that do not identify themselves as such. Furthermore, each user session is identified by a unique session ID, which is also recorded, and used in the process to identify events occurring within the same user session (i.e. if a page is reloaded by the same client within the same session, we would still count this as only one page view). Downloads are recorded for individual bitstreams (files).

The ISI Web of Science field normalized data used was provided by CWTS (The Centre for Science and Technology Studies) at Leiden University, and other citation data from ISI Web of Science has been downloaded directly from the ISI Web of Science web interface. Field normalization data was used to compensate for different citation density of diverse subject fields, publication years and publication types. As an example the field of neuroscience is cited four times more than engineering sciences (Glanzel & Moed, 2002). The field normalization score is the ratio of the actual and expected citation rate; i.e. the result after dividing the number of a publication's citations with the average number of citations worldwide (in ISI Web of Science) for this publication's subject field (as defined in ISI Web of Science's journal categories), year and publication type. Consequently, a score of 1 is indicating a world average and all scores lower than 1 mean a performance below world average (and higher than 1 above average). The field normalization score was counted on an individual paper basis, and thereafter the average was counted. For comparison we also used the "raw" citations score for the papers, based on ISI Web of Science data from the web version of the database. This data will not take into account citation density variation between different scientific fields.

The sample consisted of 158 dissertations, and the number of papers included with the comprehensive summary varies from 3 to 6. Not all papers are used for this study since they are at a manuscript level or submitted, but not yet published. The number of included papers for this study was 408, with an average of 2,7 per dissertation. All papers have been grouped for every dissertation and a mean value for the field normalization was calculated. Of all papers we detected field data for 61 % of the papers. The main reason for the "missing" articles is, of course, that the journals are not covered by ISI Web of Science. We also found some very early dated papers included for a small number of dissertations, which was not covered by the field data for our bibliometric database (which spans from 2004 an onward). The average publication date for the papers was 2007. This means that we have (as a medium) a citation window of three years for the papers, and a much shorter window for the download data. Previous research show how obsolescence patterns differ significantly between citations and downloads (Moed, 2005; Schloegl & Gorraiz, 2010), so it will not affect the results negatively using these different windows (as long as using longer windows for citations).

Correlation analysis

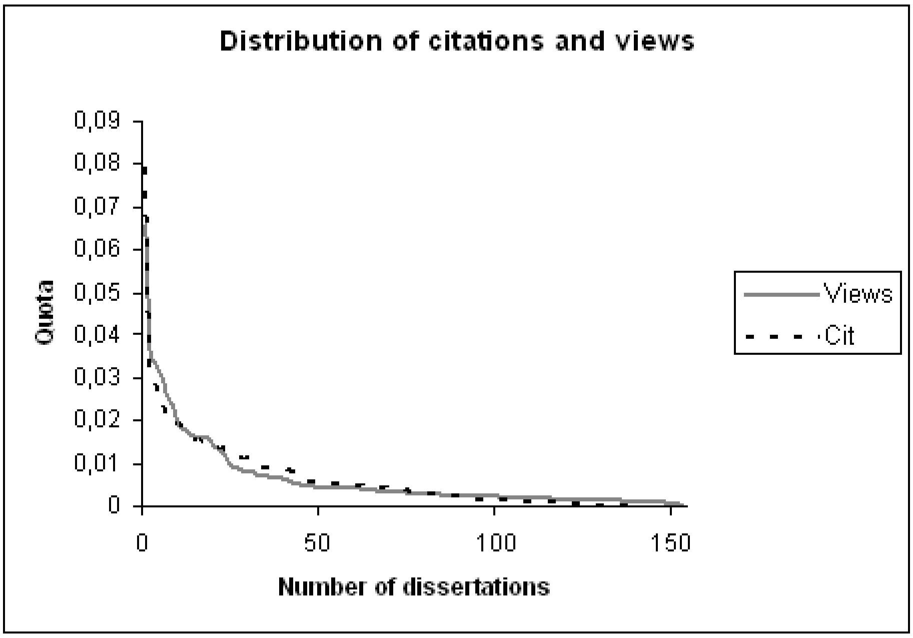
Bivariate analysis was used in order to measure the correlation between the different data and the correlation coefficient was analyzed by comparing the two matrices (the citation and download scores), using the statistics function in Excel. The coefficient is expressed by a score between -1 and 1 where 1 is a perfect positive correlation and -1 a perfect negative correlation. The score 0 means that there is no correlation at all. Another method to visualize correlations is using scatter plots. If there is a perfect positive correlation between the variables the result is a straight curve angled to the right, and a perfect negative correlation is demonstrated by a straight curve angled to the left.

Results

Distribution analysis

When working with citation data it is important to be aware of the skewed distributions of the data (Seglen, 1992). As for publication productivity, there is a “long-tail”-effect; i.e. a larger proportion of the publications receive 0 to 1 citations, and a relative small number of publications receive the majority of the citations. We examined the distribution of our data by calculating the proportional amount of downloads and citations (sum of all citations for the included papers) for every dissertation. The data was sorted cumulatively and the result demonstrates an almost identical distribution for citations and downloads. Of the 153 dissertations half of the total amount downloads was given to 21 dissertations and half of the amount citations was given to 22 dissertations. The distribution is shown in Figure 1 :

Fig 1 Similar distribution of citations and views



Lack of correlation between downloads/views and citations

There was no indication of correlation between the citation data and the views/download data. The correlation coefficient score was -0,13 for the field data and the views data, and -0,11 for the "raw" citations data (”times cited”) and the views data (see table 1 for all figures).

Table 1

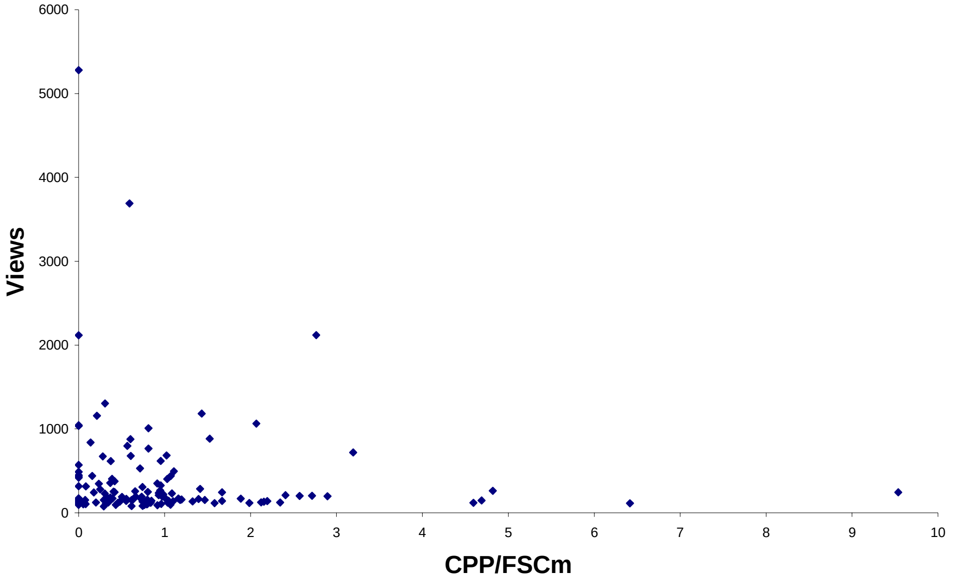
Methods compared	**Correlation score**
citation field data - views data	-0,12502
citations field data - download data	-0,04473
times cited -views data	-0,11305
times cited - download data	-0,06157

0 = no correlation
-1 = perfect negative correlation
1 = perfect positive correlation

As mentioned previously, there are some methodological problems comparing the different data. For citations we use data based on included papers and for downloads, we are using data for the summaries. This is caused by the difficulties to obtain data from commercial publishers and hosts for journal articles.

The correlation can also be visualized in a scatter plot (figure 2). As seen, the sample is clustered together, demonstrating the skewed distribution of the material. However we can also notice the lacking correlation between the variables, demonstrated by the absence of linearity between the plots.

Fig 2 Scatter plot illustrating lacking correlation



Correlation between peer judgment and citations, and PR activities and downloads/views
As shown, there was no correlation between citations and downloads. However there was a correlation between the peer judgement (the nominee group) and impact demonstrated by citations. The normalized citation score for the peer group was 1,4, which was significantly higher than the "non-nomineed" (0,88).

Another interesting relation, is the linkage between public relation activities and the volume of downloads. The average of views for dissertations with press releases was substantially higher (821) than those dissertations without releases (318). This was also indicated by a simple arithmetic operation. The twenty-three top dissertations for the different groups were compared by counting the number of times where the same dissertation appeared on the different lists. The most frequent links were those connecting PR and downloads and PR and views.

One detail observed, if not so statistically significant, is the appearance of one single dissertation on the top 10 in the views and downloads lists and top 10 in the citation lists.

This dissertation was later awarded the "best dissertation of 2009" [8]

Conclusion

This paper is the result of a praxis-oriented case study where different methods measuring usage (defined as citations and downloads) has been tested. As demonstrated, there was no correlation between the download data and citations. This could imply that the usage patterns are different between citing and downloading. However, it is important to notice the small sample of the present study, and also the different data (comprehensive summaries and papers) used when comparing. As previously described, some previous studies have shown a positive correlation between citation and download data.

Also analyzed was the relation between the dissertations promoted at the faculty website and the number of downloads. This demonstrated a positive correlation, which will strengthen the importance of the PR office, and also underline the importance of visibility when promoting science. The latter is of course important when discussing the advantages of open access publishing and e-publishing/open repository services (as the previously mentioned GUPEA). There was also a positive correlation between peer judgement and impact demonstrated by citations. This could indicate difference in interest focus from the general public (the PR activity and volume of downloads) and the scientific community (the peer judgment and citations in scientific papers).

In spite of the limited data set, this case study emphasizes the need for providing diverse data for our clients. As in the present case (“Best doctoral thesis”) it would be feasible, in the future, to provide additional data as citations. This work has also been extremely important as a way to enhance the methods used at Bibliometric and Digital services, linking citation and download data (with fairly reasonable costs and techniques). Hopefully, this could also serve as methodological inspiration for other research libraries, and also as a way to emphasize the use and value of open repository services .

References

Armbruster, C. (2008). Access, usage and citation metrics: what function for digital libraries and repositories in research evaluation? Retrieved June 29th, 2010, from http://papers.ssrn.com/sol3/Delivery.cfm/SSRN_ID1154542_code434782.pdf?abstractid=1088453

Armbruster, C. (2010). Whose metrics? Citation, usage and access metrics as scholarly information service. *Learned Publishing, 23*(1), 33-38.

Bollen, J., Van de Sompel, H., Hagberg, A., & Chute, R. (2009). A Principal Component Analysis of 39 Scientific Impact Measures. *Plos One, 4*(6), -.

Carlsson, H. (2009). Allocation of Research Funds Using Bibliometric Indicators – Asset and Challenge to Swedish Higher Education Sector. *Infotrend, 64*(4), 82-88.

Coats, A. J. S. (2005). Top of the charts: Download versus citations in the International Journal of Cardiology. *International Journal of Cardiology, 105*(2), 123-125.

[8] http://www.sahlgrenska.gu.se/aktuellt/nyheter/Nyheter+Detalj/basta-avhandlingar-utsedda-under-2009-.cid939493 (only in Swedish)

Duy, J., & Vaughan, L. (2006). Can electronic journal usage data replace citation data as a measure of journal use? An empirical examination. *Journal of Academic Librarianship, 32*(5), 512-517.

Glanzel, W., & Moed, H. F. (2002). Journal impact measures in bibliometric research. *Scientometrics, 53*(2), 171-193.

Merk, C., Scholze, F., & Windisch, N. (2009). Item-level usage statistics A review of current practices and recommendations for normalization and exchange. *Library Hi Tech, 27*(1), 151-162.

Moed, H. F. (2005). Statistical relationships between downloads and citations at the level of individual documents within a single journal. *Journal of the American Society for Information Science and Technology, 56*(10), 1088-1097.

Neylon, C., & Wu, S. (2009). Article-Level Metrics and the Evolution of Scientific Impact. *Plos Biology*, *7*(11), -.

Schloegl, C., & Gorraiz, J. (2010). Comparison of citation and usage indicators: the case of oncology journals. *Scientometrics*, *82*(3), 567-580.

Seglen, P. O. (1992). The Skewness of Science. *Journal of the American Society for Information Science*, *43*(9), 628-638.