RELATIONSHIP BETWEEN DOWNLOADS AND CITATION AND THE INFLUENCE OF LANGUAGE

Vicente P. Guerrero-Bote[1] and Félix Moya-Anegón[2]

[1] *guerrero@unex.es*
Grupo Scimago, Universidad de Extremadura, Departamento de Información y Comunicación, Plazuela Ibn Marwan, 06071 Badajoz (Spain)

[2] *felix.demoya@cchs.csic.es*
Grupo Scimago, CSIC, CCHS, IPP, C/Albasanz, 26-28, 28037 Madrid (Spain)

Abstract
Download indicators represent a great potential due to the high amount of download data that can be collected that can provide a great statistical significance. The relationship between citation and downloads at journal level and the influence of language on it is studied with the data of Scopus (for citation) and ScienceDirect (for downloads).
The results show that the use of downloads as prediction of the citation, is limited, as in the early years is when it obtained less significance. The relationship between downloads and citations is also different in different areas.
In Francophone regions the downloads of English language journals is proportionately greatly reduced with respect to their citation. There seems to be a part of the citation impact of the non-English language journals invisible in Scopus, which make the number of downloads proportionally greater than citations. This has its effect on the lack of correlation between the downloads and citations in the non-English language journals.

Conference Topic
Scientometrics Indicators: Criticism and new developments (Topic 1) and Old and New Data Sources for Scientometric Studies: Coverage, Accuracy and Reliability (Topic 2).

Introduction
The bibliometric indicators used for research evaluation not only take into account quantitative but also qualitative aspects. This is based on the citation of papers included in the main databases (Thomson Reuters Web of Science and Scopus principally) and on the idea that in spite of the different motivations (Brooks, 1985), citations are recognitions of previous works (Moed, 2005a). However, frequently, the application of these bibliometric indicators and these international databases to certain disciplines has been questioned. For some, the bibliometric indicators built from these databases are effective normally in basic science contexts in which research is spread mainly thru scientific journals (Filippo & Fernández, 2002). Different research fields have varying yearly average citation rates (Lundberg, 2007). Bibliometric indicators are almost always lower in areas of Engineering, Social Sciences and Humanities using the ISI data

(Guerrero et al., 2007) and using Scopus data (Lancho-Barrantes, Guerrero-Bote & Moya-Anegón, 2010).

Throughout the scientific literature some authors have pointed out a lack of statistical significance and normalization that could have been originated by a series of causes: the lack of database coverage in certain areas (Braun, Glänzel & Schubert, 2000, Grupo SCImago, 2006, Moya-Anegón et al., 2007), both for the journals and principally for types of documents, and the referencing habits in the different scientific areas (Broadus, 1971; Clemens et al., 1995; Cronin, Snyder & Atkins, 1997; Hargens, 2000; Kyvik, 2003; Lewison, 2001; Lindholm-Romantschuk & Warner, 1996; Nederhof et al., 1989; Nock, 2001; Price, 1970; Small & Crane, 1979; Thompson, 2002).

Since scientific literature is now mostly published and accessed online, a number of initiatives have attempted to measure scientific impact from download log data. The download data allows scientific activity to be observed immediately upon publication, rather than to wait for citations to emerge in the published literature and to be included in citation databases; a process that with average publication delays can easily take several years. Shepherd (2007) and Bollen et al, (2008) propose a Download Impact Factor as journal metrics which consists of average download rates for the articles published in a journal, similar to the citation-based JIF. Bollen et al. (2005, 2008) demonstrate the feasibility of a variety of social network metrics calculated on the basis of download networks extracted from the clickstream information contained in download log data.

Bollen et al, (2009) performed a principal component analysis of the journal rankings produced by 39 measures of scholarly impact that were calculated on the basis of both citation and download log data. Their results indicate that the notion of scientific impact is a multi-dimensional construct that cannot be adequately measured by any single indicator, although some measures are more suitable than others. They observed a greater reliability of download measures possibly caused by the high amount of download data that can be collected.

Although Kurtz et al. (2005) shows how the obsolescence function (Egghe & Rousseau, 2000) of citations and readership follow similar trajectories across time, Schloegl and Gorraiz (2010, 2011) shows that downloads and citations have different obsolescence patterns. Darmoni et al. (2002) and Bollen et al. (2009) show that journal download frequency does not correspond very much to the Impact Factor, although Schloegl and Gorraiz (2011) computed a high correlation at the journal level between citation and download frequencies when using absolute values and a moderate to high correlation when relating usage and citation impact factors. Wan, Hua, Rousseau, and Sun (2010) defined also a download immediacy index.

Although citation indicators are accepted by the international scientific community, they have problems of statistical significance and normalization that could have been originated by the lack of database coverage in certain areas, and the referencing habits in the different scientific areas.

Download indicators represent a great potential due to the high amount of download data that can be collected that can provide a great statistical significance. However, studies indicate that these are only loosely related with indicators based on impact. Would it be possible to use downloads as predictors of the citation?
And, there is no study about the influence of the language in downloads and in its relationship with citation. Are there differences between downloads and citation by language of publication? Is download number by publication languages proportional to the citation one? And does the language have influence in the origin of the citation and the downloads?
Thus our purpose is to study the relationship between citation and downloads at journal level, with the volume of data of ScienceDirect and Scopus, and the influence of language on it. The origin of download will be also studied.

Method and Data

The method that we have applied is to relate the downloads of each journal with the citation of that journal so that correlations between them could easily be found. Not to be very rough, and make a finer comparison, we have compared download counts of downloaded and downloading year with citations to cited year from citing year.
To this goal we have used the download data from ScienceDirect and the citation data from Scopus.
To set the language differences, we have studied these parameters for non-English journals in ScienceDirect. More particularly those having more than 95% of the papers in French (15), German (4) or Spanish (4) in the period 2003-2011.
We have also defined a control group of English journals in ScienceDirect, so as to establish the differences between the non-English and English journals. For every non-English journal, at least one English journal present in both databases, belonging to the same Specific Subject Area and with similar number of papers published was selected as control journal, up to 33 control journals.
To go deeper into this, we have compared the geographic origin of both the download with the citation of both groups.
Not all journals publish papers every year. There are only 8 non-English journals (French) with papers every year and 14 control journals. The rest of journals begin or are incorporated during the period, because of that they have no papers in the first years. However, there are three exceptions, one French journal with no paper the last year of the period and two control journals with no papers in the last two years.
The majority of the journals are concentrated in the Subject Area of Medicine, where all the German and Spanish journals are added and the majority of the French. Two other French journals are from "Pharmacology, Toxicology and Pharmaceutics", and three from Psychology (although one of the latter also is assigned to Medicine). The majority of the control journals are included also in “Medicine” (27) (2 in "Pharmacology, Toxicology and Pharmaceutics" and 6 in

Psychology), however, many other subject areas appear because of the addition of journals to multiple Subject Areas.

In the data from ScienceDirect supplied by Elsevier, each paper, regardless of documental type, has two dates, the online date and the publishing date. It is usual at present to publish a paper online before in the journal issue, a way to take advantage of the "Early View" effect (Moed, 2005b; Craig et al., 2007; Davis et al., 2008). We have calculated the difference between these two dates (publication date minus online date expressed in days), and it can be observed (in table 1) that in the first part of the period such difference is negative (they were published online some time after they were published in the issue). It can be observed also how during the period the difference closes to cero and becomes positive at the end of the period. That may be caused by a retrospective incorporation to ScienceDirect. For example, Masson journals started in ScienceDirect in 2010/2011, while they were already quite far in their volume numbering. E.g., Masson published volumes 1-10, and first Elsevier processed issue is volume 11 in 2010. What ScienceDirect then does is back-capture volumes 1-10 and add those to ScienceDirect. The on-line dates are then the dates the back-captured articles are added to ScienceDirect. That means that the publication/cover dates are older than the on-line dates, which gives those weird minus figures in the tables. This especially happened with Spanish titles, and also with French ones, although many French journals were already in ScienceDirect for a long time so that the effect is less. German titles from Urban&Fischer show the same patterns, although less since Elsevier acquired U&F much earlier than the Spanish publications.

The types of documents of the data provided by Elsevier ScienceDirect that accumulate more than 5% of downloads which have more than 500 downloads per paper and which accumulate a percentage of downloads superior to its percentage of papers are Review Article, Short Survey, Full length article and Short Comunication. The other types of documents do not involve major scientific contributions. Therefore in this paper we focus on these four document types from ScienceDirect as primary production.

In Scopus, documental typology is slightly different. The three types that accumulate more than 2% of the citation and more than 5 citations per paper on average are Reviews, Articles and Conference Papers. However, while this is true in this Journal Set, in general in Scopus, the Short Surveys accumulate a citation similar to the Conference Papers, so that we included it in this study, along with the above three as primary documents.

The records from ScienceDirect are 79,363, while the records from Scopus are 43,914. The divergence is mainly because Scopus covers all items except types: conference/meeting abstracts and book reviews. Specifically, the abstracts represent over 38% of the records of ScienceDirect.

Table 1: Average difference between the online date and the publication date (publication date minus online date) expressed in days. They are grouped by the year of publication in the issue.

Journal	*Language*	*Ndocc*	*2003*	*2004*	*2005*	*2006*	*2007*	*2008*	*2009*	*2010*	*2011*
Acute Pain	English	132	15.4	-68.1	21.1	37.1	43.6	37.2	54.7		
Addictive Behaviors	English	1819	190.7	106.9	177.9	235.7	224.1	140.5	129.1	127.4	122.8
Alzheimer's & Dementia	English	290			-76.0	-5.5	2.2	13.2	0.6	23.8	13.8
Asian Journal of Psychiatry	English	137						3.8	6.2	13.8	14.4
Biomedical and Environmental Sciences	English	322						-79.5	-75.4	-64.3	-66.2
Children and Youth Services Review	English	1150	6.9	71.1	145.1	233.5	111.5	178.8	153.8	139.3	138.8
Clinical Microbiology Newsletter	English	341	-95.6	-16.5	-26.5	-1.5	-1.1	-1.8	5.5	-2.2	5.5
Contraception	English	1337	-11.5	9.4	39.4	85.5	50.2	51.0	111.6	117.2	172.1
Diagnostic Microbiology and Infectious Disease	English	1789	50.9	10.4	26.7	78.5	77.9	81.4	43.0	63.1	38.4
Early Human Development	English	1014	35.3	44.8	51.7	94.4	170.9	137.8	68.3	18.4	53.0
e-SPEN, the European e-Journal of Clinical Nutrition and Metabolism	English	178					4.6	66.1	47.9	39.2	36.9
European Journal of Integrative Medicine	English	67							11.6	35.4	6.3
European Journal of Pharmaceutical Sciences	English	1368	10.5	41.4	55.4	97.0	90.0	79.3	72.5	68.3	51.2
EXPLORE: The Journal of Science and Healing	English	460			-12.3	-15.1	-24.9	-3.5	-5.4	-7.8	-4.2
Forensic Science International Supplement Series	English	30							53.6		
General Hospital Psychiatry	English	766	-17.4	-18.3	-10.5	-2.4	-4.1	22.4	98.6	102.0	48.4
International Journal of Drug Policy	English	450	1.1	49.7	23.2	35.8	142.5	214.7	256.9	195.4	62.0
International Journal of Pediatric Otorhinolaryngology Extra	English	366				51.5	75.7	123.0	228.9	359.4	309.8
Japanese Dental Science Review	English	49						33.8	54.3	152.4	139.1
Journal of Adolescent Health	English	1598	8.7	13.0	9.4	35.2	53.3	84.5	114.8	121.4	139.5
Journal of Cardiology Cases	English	175								127.9	70.8
Journal of Hepatology	English	2325	35.3	69.4	90.9	99.3	97.3	76.2	81.0	84.6	193.4
Journal of Medical Colleges of PLA	English	259					-90.6	-39.3	-70.7	-64.5	-64.5
Journal of Pediatric Urology	English	670			117.9	197.2	200.5	155.6	143.3	212.3	199.6
Journal of the American Academy of Child & Adolescent Psychiatry	English	1278	-2327.0	-1970.0	-1595.0	-1238.6	-864.5	-466.4	-94.0	33.7	32.8
Mental Health and Physical Activity	English	50						38.2	109.5	103.2	122.4
Microbial Pathogenesis	English	711	22.9	42.2	13.2	35.7	84.0	118.2	70.4	100.3	95.5
Nanomedicine: Nanotechnology, Biology and Medicine	English	403			-53.1	-13.8	5.5	76.4	182.5	198.9	194.4
Progress in Lipid Research	English	208	80.9	87.8	3.7	80.3	80.4	109.2	76.2	144.4	105.0
Research in Social and Administrative Pharmacy	English	228			-26.1	-8.5	-19.4	25.5	143.2	213.7	331.3
Sexologies	English	222				-32.3	16.5	76.2	38.4	88.9	83.5
Surgical Pathology Clinics	English	131						-5.0	-12.5	-58.6	-4.5
Taiwanese Journal of Obstetrics and Gynecology	English	598		-1812.7	-1460.6	-1089.9	-415.2	-109.1	-48.3	-54.2	-43.1
Actualités Pharmaceutiques	French	461						-238.7	-74.2	-67.0	-112.0
Actualités Pharmaceutiques Hospitalières	French	183			-1290.0	-952.6	-577.7	-260.0	-85.6	-153.6	-135.4
Annales Médico-psychologiques, revue psychiatrique	French	957	-29.3	20.6	37.1	85.5	104.0	224.9	85.9	55.2	81.0
Archives de Pédiatrie	French	2732	-102.0	26.4	45.5	33.6	17.8	-7.3	24.7	-4.5	10.3
Gynécologie Obstétrique & Fertilité	French	1206	-53.3	19.3	4.5	-0.8	-41.8	-3.0	-2.2	2.7	14.0
Journal de Pédiatrie et de Puériculture	French	367	-130.2	54.9	30.4	31.5	15.9	22.4	29.6	48.8	59.0
La Revue de Médecine Interne	French	1573	-248.9	-20.8	45.6	81.5	87.0	139.7	144.8	83.2	151.0
La Revue de Médecine Légale	French	29								-16.8	8.6
L'Encéphale	French	968		-1239.1	-862.3	-428.0	-127.9	46.5	88.2	108.5	87.4
Médecine & Droit	French	224	-147.5	-80.8	-42.2	-3.5	22.1	4.0	16.9	14.9	37.8
Neuropsychiatrie de l'Enfance et de l'Adolescence	French	613	-36.2	-24.3	-43.0	-6.5	16.8	66.4	112.4	173.7	201.1
Nutrition Clinique et Métabolisme	French	273	-18.6	9.1	17.0	-62.1	-58.1	-15.9	2.0	9.7	7.8
Pratiques Psychologiques	French	244		9.3	14.9	36.2	53.8	62.1	245.7	329.9	472.2
Psychologie Française	French	202		14.6	81.6	80.0	145.8	115.9	111.0	49.0	33.8
Réanimation	French	132	-25.3	-2.9	-14.1	-6.7	-74.4	-39.5	65.5	-35.7	
Das Neurophysiologie-Labor	German	63					-2.6	5.6	153.5	58.6	59.5
Krankenhaus-Hygiene + Infektionsverhütung	German	105					-5.1	19.9	1.0	30.7	21.0
Osteopathische Medizin	German	70						-151.1	-58.2	-62.3	-30.4
Public Health Forum	German	282					23.6	10.4	44.8	38.1	38.2
Cardiocore	Spanish	90								-13.0	83.6
Revista de Psiquiatría y Salud Mental	Spanish	69						-20.0	-116.0	-62.0	-41.2
Revista Española de Patología	Spanish	243				-1448.7	-1081.4	-706.7	-348.1	-20.8	61.3
Revista Internacional de Acupuntura	Spanish	171					-416.8	-197.1	-68.9	-113.4	-147.5

It may seem a bit inconsistent that one documental type from Scopus considered primary production are "Conference Papers", while the type "Conference" from ScienceDirect has not been considered. However, the percentage involved is quite small, and the percentage of downloads which accumulates is even lower, which means that the number of downloads per papers is below average. And at the same time the "Conference Papers" from Scopus are included primarily as "Full

Length Articles" and only less than 5% of them are listed as "Conference" from ScienceDirect, because ScienceDirect assign 'Full Length Article' to full scientific papers in Conference issues. Then, though "Conference Papers" of Scopus represents the greater part of "Conference" of ScienceDirect, they do not get a large number of downloads.

Results and discussion

In Figure 1, the average of citation from primary documents considered have been represented per Scopus documental types and age. Unlike other similar representations in this case they have not been made per calendar year, but the time difference between the citing and cited document has been considered. That is, for instance to calculate the citation average in the eighth year, only the papers with a minimum of 8 years have been considered, and the average was calculated with citation aged between 7 and 8 years. As in Scopus only pre-2012 citation can be considered almost complete, to calculate the citation average of 8 years only papers published in 2003 were considered since they are the only ones having at least eight full years to receive citations. To calculate the citation average of seven years papers published in 2003 and 2004 were considered since they are the only ones having at least 7 full years to receive citations. And so on. This means that the data for lesser age are statistically more significant because they have been computed with larger datasets.

The citation maximum for Reviews is obtained at 3 years. For articles, the citation for the third and fourth years is very similar, and the fall is much slower, the citation in the seventh year exceeds even that of the second year.

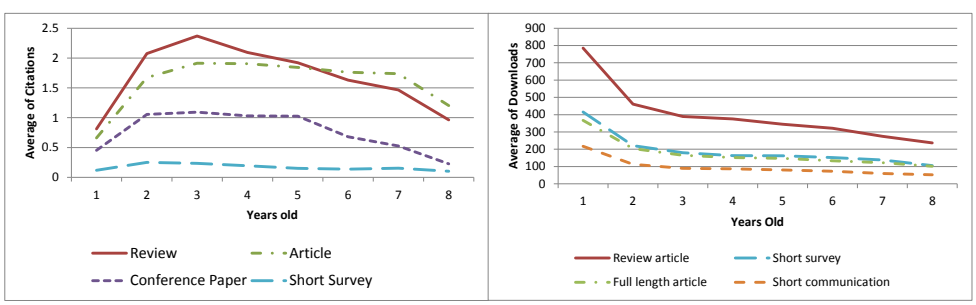

Figure 1: Average of primary Citations per Scopus document type by age in years of the Journal Set and Averages of Downloads of the main Science Direct document types per Science Direct paper by years of difference with respect to the online publication date.

In the second part of Figure 1, a similar representation has been made but with regard to the downloads. In this case, the online date has been used as reference. Similarly, the number of years shown on the horizontal axis refers to the difference between the date of download and online publication date. To calculate each average, only those papers that have the corresponding full annuity to be

downloaded have been considered. In 8, only those downloads in which the difference between the download and the online date is between 7 and 8 years have been considered.
In this case all the curves are monotonic decreasing, while in the previous figure as a result of the time required for citation from the date in which the cited paper is published until the citing paper is made and published.
In the case of downloads, the diffusion made when the paper is published online and the large number of downloads that results from novelty are very evident. Also there is a greater difference between Reviews and Articles.
Figure 2 shows the citation from primary papers toward primary papers by Subject Areas. The way of computation is similar to the previous one. The peak in the seventh year of Pharmacology, Toxicology and Pharmaceutics is striking, however it is because in 2005 two of the four journals of the subject area enter, which lower significantly the citation average, except for that incidence the curves are quite similar.
The second part of Figure 2 similarly shows download averages by Subject areas and years. As in the first part, irregularities in the curve of Pharmacology, Toxicology and Pharmaceutics are observed, following the incorporation of the four journals assigned to it at different times.
Also striking is the difference in the order of the Subject Areas, while Medicine is the one with the highest average of citations, it is the one that has the lowest downloads. Psychology while always behind in citations, is always ahead in downloads. This may, once again, be indicative of different patterns in different areas.

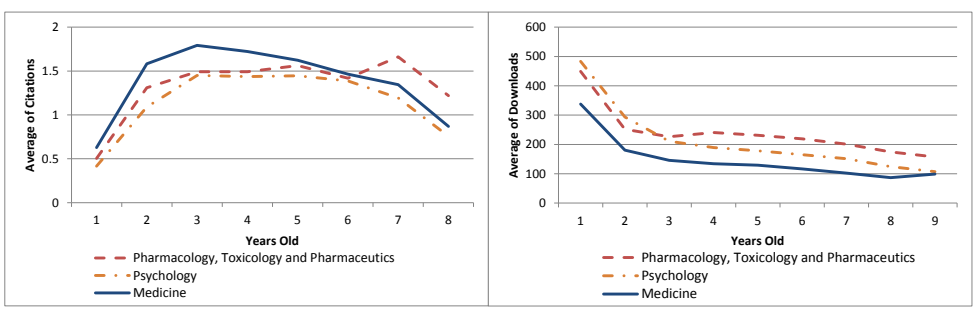

Figure 2: Average of primary citation toward primary papers, and average of downloads of primary papers by Subject Area (only the three original Subject Areas).

Table 2 shows the correlation between averages of downloads and averages of citations by journals, year of publication and of citation or download "age" using the same calculation method as above. The only difference is that, in order to allow both data to be comparable, for download "age", the date of publication of the paper in the journal has been used instead of online publication date as above. In columns the age of the citation and in rows the age of the downloads. The last

column and row correspond to the sum of all averages of citation/downloads as an average of citation/downloads of up to 8 years old. There is a certain time delay between download and citation: if an author downloads an article, he must first read it, include it a new paper he is writing, and that paper must be published. All this may take 1-2 years, sometimes even more, depending upon journal and perhaps field.

Table 2: Correlations between averages of downloads and averages of citations by journals, year of publication and of citation or download "age". In columns the age of the citation and in rows the age of the downloads. The last column/row correspond to the sum of all averages of citation/downloads.

DY\CY	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*Σ*
1	0.77	0.78	0.82	0.85	0.86	0.88	0.93	0.94	0.51
2	0.71	0.75	0.79	0.84	0.87	0.89	0.93	0.94	0.6
3	0.66	0.71	0.76	0.79	0.83	0.85	0.92	0.93	0.63
4	0.63	0.69	0.74	0.77	0.78	0.81	0.86	0.92	0.67
5	0.64	0.68	0.73	0.75	0.76	0.75	0.85	0.89	0.7
6	0.61	0.66	0.7	0.72	0.75	0.76	0.81	0.9	0.71
7	0.73	0.73	0.76	0.77	0.8	0.81	0.82	0.77	0.78
8	0.72	0.72	0.74	0.78	0.8	0.8	0.84	0.82	0.79
Σ	0.65	0.71	0.76	0.77	0.79	0.79	0.83	0.82	0.73

Table 2 shows that the highest correlation is between the number of downloads of one year of difference and the citation 8 years of difference. The average correlation between the number of downloads and the citation of the same age is 0.78 (the diagonal), between the number of downloads and the citation of two more years of difference is 0.84 and between the citation and the number of downloads of 2 years more of difference 0.73. These results are consistent with the idea that there is a certain time delay between download and citation.

Table 3 shows the correlations between downloads and citations of two more years of age separated by groups with levels of statistical significance. The correlations are significant and positive for the total and for the control set. However, they are lower (in some cases even they may be slightly negative) and of little significance in the case of non-English language journals.

Table 4 has been made with the same data as table 3, but in this case columns of downloads have been correlated with the column that sums the averages of citation up to 8 years of age. With this you can see which are the most significant downloads when predicting total citations obtained by each journal. In this case we can see that none of the correlations of the non-English language journals are statistically significant at a level of $\alpha = 0.05$. The most significant is the third year with a correlation below 0.2.

Table 3: Correlations between averages of downloads and averages of citations by journals, year of publication and of citation or download "age". The citation "age" is two years more than download "age". The correlations have been separated by language of publication.

		1->3	*2->4*	*3->5*	*4->6*	*5->7*	*6->8*	*Σ*
Total	*r*	0.82	0.84	0.83	0.81	0.85	0.90	0.73
	α	<0.01	<0.01	<0.01	<0.01	<0.01	<0.01	<0.01
English	*r*	0.80	0.86	0.83	0.82	0.87	0.90	0.71
	α	<0.01	<0.01	<0.01	<0.01	<0.01	<0.01	<0.01
Non-English	*r*	0.29	0.33	0.21	-0.03	-0.31	0.36	0.43
	α	0.01	0.02	0.16	0.89	0.18	0.34	<0.01

Table 4: Correlations between averages of downloads and the sum of the averages of citations by journals, year of publication and of citation or download "age". The correlations have been separated by language of publication.

		1	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*Σ*
Total	*r*	0.51	0.60	0.63	0.67	0.70	0.71	0.78	0.79	0.73
	α	<0.01	<0.01	<0.01	<0.01	<0.01	<0.01	<0.01	<0.01	<0.01
English	*r*	0.45	0.56	0.60	0.64	0.68	0.69	0.77	0.74	0.71
	α	<0.01	<0.01	<0.01	<0.01	<0.01	<0.01	<0.01	<0.01	<0.01
Non-English	*r*	0.07	0.15	0.20	0.12	0.04	-0.005	-0.09	0.27	0.43
	α	0.46	0.13	0.08	0.34	0.78	0.98	0.71	0.48	<0.01

The correlations of the control journals are statistically significant and positive. The highest correlation is obtained in the seventh year.

However, in many cases, only the citation obtained in the first three years is taken into account, which is why we have created Table 5 which shows the correlation of the averages of downloads of different ages with the sum of the averages of citation of the first three years. In this case the correlation increases slightly in the early years in cases of non-English journals although the correlation is still quite low. In the case of the control journals, all the correlations have a high level of statistical significance, and the maximum value is obtained in the downloads of seven years of "age", although the first three years rise steeply with respect to the correlations in Table 4.

For the study and comparison of the origin of the downloads and the citation, two tables were generated by countries with the number of citations and downloads of each country to the control journals of English language in one column, to the French-language journals in another, to the German language journals in another and to the Spanish language journal in the fourth. We have also calculated other columns with the total number of citations and downloads and with citations and downloads of the three groups of journals to study (the French language journals,

the German language journals and Spanish language journals). The countries in this table have been ordered by the scientific production in the period. From them we have kept the data of the 50 most productive countries, which are those with more than 25,000 papers in the period 2003-2011.

Table 5: Correlations between averages of downloads and the sum of the averages of citations of up to 3 years old by journal, year of publication and of citation or download "age". The correlations have been separated by language of publication.

		1	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*Σ*
Total	*r*	0.66	0.74	0.73	0.72	0.70	0.68	0.75	0.74	0.75
	α	<0.01	<0.01	<0.01	<0.01	<0.01	<0.01	<0.01	<0.01	<0.01
English	*r*	0.61	0.72	0.71	0.70	0.68	0.65	0.73	0.67	0.73
	α	<0.01	<0.01	<0.01	<0.01	<0.01	<0.01	<0.01	<0.01	<0.01
Non-English	*r*	0.24	0.33	0.33	0.20	0.12	0.09	-0.01	0.27	0.54
	α	<0.01	<0.01	<0.01	0.12	0.44	0.63	0.98	0.49	<0.01

By correlating the columns, correlations higher than 0.98, statistically significant at $\alpha = 0.01$ level were found between downloads and citations to the same type of journals. The downloads and citations of the control journals have correlations higher than 0.93, statistically significant at $\alpha = 0.01$ level with the scientific production and the total sum of downloads and citations (to the control journals and non-English language journals) while downloads and citations of non-English language journals do not correlate significantly, either by language or as a set (only two correlations are significant at $\alpha = 0.01$, which are slightly less than 0.5, both with the total sum of downloads in each country, one of them is of the citations to French journals and the other of the citations to non-English language journals).

As Table 6 shows, the countries with the highest percentage of downloads of control journals (relative to the total number of downloads thereof) are USA, China and UK. The following country is Canada, although it has a higher percentage of downloads of the French journals. The countries with the highest percentage of the French Journals downloads are France, Tunisia, Canada, Algeria and Belgium all Francophones, although Tunisia and Algeria are not among the 50 most productive. Germany, Switzerland and Austria are for German Journals. Switzerland also has a high percentage of downloads from the French Journals. Spain, Brazil, Argentina and Uruguay are for Spanish Journals, the last two at a great distance, and Brazil not being a Spanish speaking country. If we compare these percentages of downloads with the percentage of total downloads, we see that the three countries with the highest percentage of Control Journals have a slightly higher percentage of them than the total, the ratio is slightly greater than unity. In the case of the French Journals, the ratio becomes slightly larger. Curiously, Tunisia and Algeria have the highest ratio. This ratio continues

to increase in the case of the German Journals and especially the Spanish Journals indicating that these downloads are more concentrated in those countries.

Table 6: The Highest Percentages of Downloads of the Control (%CD), French (%FD), German (%GD) and Spanish (%SD) journal group with respect to the total number of downloads of each group, ratio of these percentages with respect to the Download percentage of each country (%TD) and similar citation ratios.

Country	*%CD*	*%FD*	*%GD*	*%SD*	*%CD/%TD*	*%FD/%TD*	*%GD/%TD*	*%SD/%TD*	*%CC/%TC*	*%FC/%TC*	*%GC/%TC*	*%SC/%TC*
United States	34.41%	4.86%	4.74%	0.46%	1.242	0.175	0.171	0.017	1.060	0.229	0.019	0.119
United Kingdom	8.66%	1.72%	2.24%	0.17%	1.222	0.243	0.316	0.024	1.041	0.474	0.186	0.197
China	6.13%	1.83%	4.68%	0.34%	1.188	0.354	0.906	0.066	1.058	0.256	0	0
France	2.70%	48.08%	1.06%	0.01%	0.211	3.752	0.083	0.001	0.599	6.305	0.091	0
Tunisia	0.27%	10.63%	0.10%	0.01%	0.106	4.115	0.038	0.004	0.331	9.842	0	0
Canada	4.53%	5.49%	0.47%	0.14%	0.958	1.161	0.098	0.029	1.029	0.634	0.430	0
Algeria	0.10%	4.31%	0.05%	0.00%	0.093	4.161	0.052	0	0.539	7.107	0	0
Belgium	0.90%	3.25%	0.86%	0.00%	0.633	2.282	0.608	0	0.916	2.118	0	0.299
Germany	1.91%	0.50%	55.72%	0.04%	1.107	0.288	32.320	0.025	1.029	0.532	14.633	0
Switzerland	1.66%	3.04%	6.50%	0.03%	0.840	1.535	3.287	0.017	0.972	1.380	0.674	0
Austria	0.24%	0.03%	5.40%	0.00%	1.169	0.142	26.159	0	1.028	0.621	3.088	0
Spain	1.80%	1.77%	0.51%	78.59%	0.956	0.939	0.269	41.707	0.979	1.110	0	17.820
Brazil	2.02%	1.15%	0.40%	16.23%	1.098	0.626	0.217	8.803	1.008	0.903	0	1.428
Argentina	0.20%	0.04%	0.01%	1.44%	1.208	0.246	0.035	8.712	1.012	0.844	0	1.196
Uruguay	0.01%	0.02%	0.00%	0.92%	0.749	1.555	0.152	60.521	0.938	1.839	0	0

If you calculate a similar ratio to citation, we can see that the control journals have lower ratios (that of downloads) in the case of USA, UK and China and higher in the remaining cases. The French Journals have higher citation ratios in most of the countries shown. The German Journals only in France and Canada. The Spanish Journals only in the case of USA, UK and Belgium.

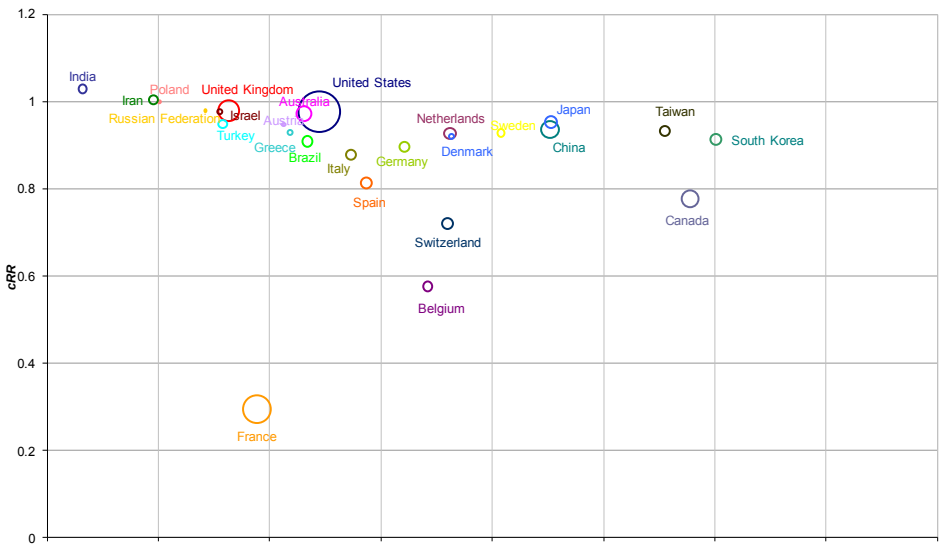

Figure 3: Ratio of downloads with respect to citations of the journals of control (cRR) against the French, German and Spanish language journals (eRR). The size is proportional to the number of total national downloads. The 27 countries with the highest scientific production are represented.

Some others relative columns have also been calculated per country, as the ratio of downloads with respect to citations:

$$gRR_{country} = \frac{\frac{gd_{country}}{td_{country}}}{\frac{gc_{country}}{tc_{country}}}$$

Where *d* are downloads, of a journal group (*gd*) or in total (*td*), and *c* are citations to a journal group (*gc*) or in total (*tc*). The measurements calculated in this way for each country were: *cRR* (ratio of downloads of the control journals with respect to its citations ratio), *fRR* (ratio of downloads of the French journals with respect to its citations ratio), *gRR* (ratio of downloads of the German journals with respect to its citations ratio), *sRR* (ratio of downloads of the Spanish journals with respect to its citations ratio), *eRR* (ratio of downloads of the French, German and Spanish journals with respect to its citations ratio).

Regarding download ratios with respect to the citations, we found that the control journals (*cRR*) average in the top 50 countries of 0.93 with a standard deviation of 0.12. While the group consisting of French, German and Spanish language journals (*eRR*) has an average of 2.35 with a standard deviation of 1.36 (mean difference significant at the $\alpha = 0.01$). This means that in these countries the control journals are cited in greater proportion to the downloading, while the French, German and Spanish languages journals are downloaded twice with respect to its citation. This can be seen in figure 3.
As you can see countries with the lowest downloaded papers of the control journals with respect to the citing, have a francophone link. This effect does not occur in the case of Spanish or German, but we must take into account that the German and Spanish journals studied are very few, and some of them seem to have been included on ScienceDirect retrospectively.

Conclusions

The number of papers from Scopus and ScienceDirect is different, because the first includes all items except types: conference/meeting abstracts and book reviews. The divergence is mainly because of conference/meeting abstracts.
The set of journals in German and Spanish language is not very significant in order to find separate conclusions.
The citation and download curves with respect to time are different. The time required for a paper to be cited can be seen in the citation curves and the effect of novelty in download curves. The proportional difference between the downloads received by the reviews and other types of documents increases with respect to the citation.bi
The order of the Subject Areas in average citation does not match the order in average download. This leads to different patterns in different areas i.e. researchers in different areas cite proportionally differently with respect to what they read.
There are statistically significant correlations between the downloads and citations for journals and years, but these are greatly reduced in both value and statistical significance in the case of non-English language journals. Some influence on these results can have the late incorporation of these journals to ScienceDirect.
In the control journals, at first there is a novelty effect that makes many downloads occur that do not result in citations. This may be the reason why the first year is the one which obtain lower correlations. Interestingly the highest correlations are those of the sixth or seventh year of age, which may correspond to when researchers are looking for a particular paper probably redirected by a citation.
All this makes the use of downloads as prediction of the citation, limited, as in the early years is when it obtained less significance. In no case thus does it reach the correlation between the citation of the first three years with the citation total

(0.91). This circumstance is even greater in the case of non-English language journals.
The 50 most productive countries download the control journals proportional slightly less than they cited them. However the non-English journals to study are downloaded proportionately more than twice what they are cited. This may be due to the fact that a part of the citation impact of the non-English journals is invisible in Scopus because those who download the papers, also cite them in articles published in journals that are not processed for Scopus.
In these 50 most productive countries, there is an association between the proportional citation or downloads of control journals with the ratio between downloads and citation of them. This means that those which frequently proportionally download or cited the control journals, download them proportionately more with respect what they cite them. This same effect does not occur in the non-English journals to study.
In Francophone regions it is observed how the download of control journals is proportionately greatly reduced with respect to their citation. In the case of German and Spanish language, the study is not very significant because the number of journals is very small, some of which have been loaded into ScienceDirect retrospectively.
Definitely there seems to be a part of the citation impact of the non-English language journals invisible in Scopus, which make the number of downloads proportionally greater than citations. This has its effect on the lack of correlation between the downloads and citations in the non-English journals, which means that the downloads can hardly be used to predict the citation.

Acknowledgments

This work was granted by Elsevier as part of the Elsevier Bibliometric Research Program (EBRP) and financed by the Junta de Extremadura, Consejería de Empleo, Empresa e Innovación and the Fondo Social Europeo as part of the research group grant GR10019.

References

Bollen, J., Van de Sompel, H. and Rodriguez, M.A. (2008). Towards usage-based impact metrics: First results from the MESUR project. In *Joint Conference on Digital Libraries (JCDL2006)*, Pittsburgh, PA, June 2008.

Bollen, J., Van de Sompel, H., Hagberg, A, and Chute, R. (2009). A principal component analysis of 39 scientific impact measures. *PLoS ONE*, 4(6): e6022. doi:10.1371/journal.pone.0006022.

Bollen, J., Van de Sompel, H., Smith, J. and Luce, R. (2005). Toward alternative metrics of journal impact: a comparison of download and citation data. *Information Processing and Management*, 41(6):1419-1440.

Braun, T., Glänzel, W., Schubert, A. (2000). How balanced is the Science Citation Index's journal coverage? A preliminary overview of macrolevel statistical data. In: Cronin, B; Barsky Atkins, H (eds.). *The Web of knowledge,*

a festschrift in honor of Eugene Garfield. Canada: American Society of Information Science, 2000, pp. 251–277.
Broadus, R. N. (1971). The literature of the social sciences: a survey of citation studies. *International Social Sciences Journal*, 23: 236–243.
Brooks, T.A. (1985). Private acts and public objects: an investigation of citer motivations. *Journal of the American Society for Information Science*, 36(4): 223-229.
Clemens, E. S., Powell, W. W., Mcilwaine, K., Okamoto, D. (1995). Careers in print: Books, journals, and scholarly reputations. *American Journal of Sociology*, 101: 433–494.
Craig, I., Plume, A., McVeigh, M., Pringle, J., Amin, M. (2007). Do open access articles have greater citation impact? A critical review of the literature. *Journal of Informetrics*, 1, 239-48.
Cronin, B., Snyder, H., Atkins, H. (1997). Comparative citation rankings of authors in monographic and journal literature: a study of sociology. *Journal of Documentation*, 53: 263–273.
Darmoni, S. J., Roussel, F., Benichou, J., Faure, G. C., Thirion, B., & Pinhas, N. (2000). Reading factor as a credible alternative to impact factor: a preliminary study. *Technol. Health Care*, 8 (3-4), 174–175.
Davis, Philip M., Bruce V. Lewenstein, Daniel H. Simon, James G. Booth, and Matthew Connolly. (2008). Open Access Publishing, Article Downloads, and Citations: Randomized Controlled Trial. *British Medical Journal*, 337: 331-345.
Egghe, L., & Rousseau, R. (2000). Aging, obsolescence, impact, growth, and utilization: Definitions and relations. *Journal of the American Society for Information Science*, 51 (11), 1004–1017.
Filippo, D., Fernández, M.T. (2002). Bibliometría: importancia de los indicadores bibliométricos. In: *El estado de la ciencia*. p. 69-76. Red Iberoamericana de Indicadores de Ciencia y Tecnología (RICYT).
Grupo SCImago (2006). Análisis de la cobertura de la base de datos Scopus. *El profesional de la información*, Vol.15, nº2: 144-145.
Guerrero-Bote, V. P., Zapico-Alonso, F., Espinosa-Calvo, M. E., Gómez-Crisóstomo, R., & Moya-Anegón, F. (2007). The Iceberg Hypothesis: Import-Export of Knowledge between scientific subject categories. *Scientometrics*, 71(3): 423-441.
Hargens, L. L. (2000). Using the literature: reference networks, reference contexts, and the social structure of scholarship. *American Sociological Review*, 65 : 846–865.
Kurtz, M. J., Eichhorn, G., Accomazzi, A., Grant, C. S., Demleitner, M., & Murray, S. S. (2005). The bibliometric properties of article readership information. *Journal of the American Society for Information Science and Technology*, 56: 111-28.
Kyvik, S.(2003). Changing trends in publishing behaviour among university faculty, 1980–2000. *Scientometrics*, 58: 35–48.

Lancho-Barrantes, B.S., Guerrero-Bote, V.P. , Moya-Anegón, F. (2010). The Iceberg Hypothesis revisited. *Scientometrics*, 85: 443-461.
Lewison, G. (2001), Evaluation of books as research outputs in history of medicine. *Research Evaluation*, 10 : 89–95.
Lindholm-Romantschuk, Y., Warner, J. (1996). The role of monographs in scholarly communication: an empirical study of philosophy, sociology and economics. *Journal of Documentation*, 54 : 389–404.
Lundberg, J. (2007). Lifting the crown—citation z-score. *Journal of Informetrics*, 1, 145–154 .
Moed, H.F. (2005a). *Citation Analysis in research evaluation*. Dordrecht; Springer, p. 346.
Moed, H.F. (2005b). 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, 1088-97.
Moya-Anegón, F., Chinchilla-Rodríguez, Z., Vargas-Quesada, B., Corera-Álvarez, E., Muñoz-Fernández, F. J., González-Molina, A., et al. (2007). Coverage analysis of Scopus: a journal metric approach. *Scientometrics*, 73, (1) , 53-78.
Nederhof, A. J., Zwaan, R. A., De Bruin, R. E., Dekker, P. J. (1989). Assessing the usefulness of bibliometric indicators for the humanities and the social sciences. *Scientometrics*, 15 : 423–435.
Nock, D. A. (2001). Careers in print: Canadian sociological books and their wider impact, 1975–1992. *Canadian Journal of Sociology/Cahiers canadiens de sociologie*, 26: 469–485.
Price, D. J. (1970). Citation measures of hard science, soft science, technology, and non-science. In: C. E. Nelson, D. Pollack (Eds), *Communication Among Scientists and Engineers*. Lexington, Mass., Lexington books.
Schloegl, C., & Gorraiz, J. (2010). Comparison of citation and usage indicators: The case of oncology journals. *Scientometrics*, 82(3), 567–580.
Schloegl, C., & Gorraiz, J. (2011). Global Usage Versus Global Citation Metrics: The Case of Pharmacology Journals. *Journal of the American Society for Information Science and Technology*, 62(1):161–170.
Shepherd, P.T. (2007). The feasibility of developing and implementing journal usage factors: a research project sponsored by UKSG. *Serials: The Journal for the Serials Community*, 20(2):117-123.
Small, H. G., Crane, D. (1979). Specialties and disciplines in science and social science: an examination of their structure using citation indexes. *Scientometrics*, 1 : 445–461.
Thompson, J. W. (2002). The death of the scholarly monograph in the humanities? Citation patterns in literary scholarship. *Libri*, 52 (3) : 121–136.
Wan, J.-K., Hua, P.-H., Rousseau, R., & Sun, X.-K. (2010). The journal download immediacy index (DII): Experiences using a Chinese full-text database. *Scientometrics*, 82(3), 555–566.