Download vs. citation vs. readership data: the case of an information systems journal

Christian Schlögl[1], Juan Gorraiz[2], Christian Gumpenberger[2], Kris Jack[3] and Peter Kraker[4]

[1]*christian.schloegl@uni-graz.at*
University of Graz, Institute of Information Science and Information Systems, Universitätsstr. 15/F3, A-8010 Graz (Austria)

[2]*(juan.gorraiz|christian.gumpenberger)@univie.ac.at*
University of Vienna, Vienna University Library, Dept of Bibliometrics, Boltzmanngasse 5, A-1090 Vienna (Austria)

[3] *kris.jack@mendeley.com*
Mendeley, 144a Clerkenwell Road, London, EC1R5DF (UK)

[4]*pkraker@know-center.at*
Know-Center, Inffeldgasse 13, A-8010 Graz (Austria)

Abstract
In our article we compare downloads from ScienceDirect, citations from Scopus and readership data from the social reference management system Mendeley for articles from the Journal of Strategic Information Systems (publication years: 2002-2011). Our study shows a medium to high correlation between downloads and readership data (Spearman r=0.73) and between downloads and citations (Spearman r=0.77). However, there is only a medium-sized correlation between readership data and citations (Spearman r=0.51). These results suggest that there is at least "some" difference among the two usage measures and the (citation) impact of the analysed information systems articles. As expected downloads and citations have different obsolescence characteristics. While the highest downloads accrue the first years after publication, it takes several years until the citation maximum is reached.

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

Introduction
There exist already a slew of studies which have compared download and citation data. These studies can be divided to two groups: investigations having been performed at local level and those having been conducted at global level (Bollen and van de Sompel, 2008). While the former are restricted to a specific user population (e.g. a university), global studies are performed on a world-wide context. Usually they use download data from repositories/pre-print archives, open access journals or e-journals from (commercial) publishers as primary data source. Examples for the latter can be found, for instance, in Moed (2005) and in Schloegl and Gorraiz (2010, 2011).

With the advent of the social web and its growing acceptance in academia, alternative metrics seem to be a further source for the measurement of science (Bar-Ilan et al, 2012). In particular, what is called "longer-term metrics" in an editorial of a Nature article (Anonymous, 2012) seems promising. These metrics are based on downloads, readers and user comments. An example is the social reference management system Mendeley. So far social media has not been accepted as part of the measurement of scientific achievement because it has not yet been sufficiently validated. The few investigations which are known to the authors can be found in Bar-Ilan (2012), Bar-Ilan et al. (2012), Kraker et al. (2012), Li, Thelwall and Giustini (2012) and Li and Thelwall (2012). As a consequence, this research in progress paper is to

provide one more evidence concerning the potential of social media using the example of Mendeley. In particular, the following issues will be addressed:

- Are most cited articles the most downloaded ones and those which can be found most frequently in user libraries of the collaborative reference management system Mendeley?
- Do citations and downloads have different obsolescence characteristics at publication level?
- Are there other features in which citation, download and readership data differ?

Methodology and data sources

All the following analyses were performed for the Journal of Strategic Information Systems. Both citations and downloads were provided by Elsevier in the framework of the Elsevier Bibliometric Research Program (EBRP). For all documents published between 2002 and 2011 all monthly downloads were made available from ScienceDirect and all monthly citations from Scopus until mid of 2012. Furthermore, we received the total number of occurrences of full length articles in user libraries in Mendeley from 2002 to 2011.

Mendeley provides users with software tools that support them in conducting research (Henning & Reichelt 2008). One of the most popular of these tools is Mendeley Desktop, a cross-platform, freely downloadable PDF and reference management application. It helps users to organize their personal research libraries by storing them in relevant folders and applying tags to them for later retrieval. The articles, provided by users around the world, are then crowd-sourced into a single collection called the Mendeley research catalogue (see Hammerton et al. (2012) for details). At the time of writing, this catalog contains more than 80 million unique articles, crowd-sourced from over 2 million users, making it an interesting source of data for large scale network analysis.
Furthermore, Mendeley enables users to create and maintain a user profile that includes their discipline, research interests, biographical information, contact details, and their own publications. Mendeley then takes this data and automatically generates a profile page for the user that acts as a CV in which they can showcase their expertise. The user's publications are also augmented by readership counts, allowing them to track the popularity of their individual papers within the Mendeley community. These readership counts indicate how many Mendeley users have added the author's article to their personal research library.
To find corresponding articles in the Mendeley catalog, we matched paper titles reported from Elsevier to the titles of articles in the Mendeley database. Since there can be slight differences between article title across the two databases, we employed the Levenshtein distance when matching them up to one another in order to take account of these inconsistencies. We found good matching results of around 99.9% accuracy when employing a Levenshtein ratio of 1/15.83. Nevertheless, we manually verified borderline cases to reduce the likelihood of false positive matches.

Results

Download data

Table 1. Downloads per download type (pdf or HTML) (publication years: 2002-2011, n=321 docs, download years: <=2011)

Download type	*%*
HTML	39%
Pdf	61%

There are two download types available in ScienceDirect from which pdf was used most (approximately in 61% of all cases between 2002 and 2011 – see Table 1) for the information systems journal under consideration.
As can be seen in Table 2, 94 per cent of all downloads allotted to full length articles (FLAs) which have a proportion of 56 per cent among all document types in ScienceDirect. As a consequence, the number of downloads per document is by far the highest for this document type. Interestingly, documents of other types are also downloaded to some extent, even though several magnitudes lower.

Table 2. Distribution of document types (n=321 documents) and downloads (publication year: 2002-2011, download year: <=2011) per document type.

Document type	*n*	*% docs*	*% downloads*	*Downloads per doc – relations*[1]
Announcement	5	1.6%	0.4%	5.9
Book review	4	1.2%	0.3%	5.5
Contents list	29	9.0%	0.4%	1.0
Editorial Board	29	9.0%	0.6%	1.5
Editorial	49	15.3%	3.3%	4.6
Erratum	1	0.3%	0.1%	5.7
Full length article	181	**56.4%**	**94.1%**	**35.4**
Index	12	3.7%	0.2%	1.3
Miscellaneous	9	2.8%	0.2%	1.8
Publishers note	2	0.6%	0.2%	7.0
	321	100%	100%	

[1] Since the download numbers are very sensitive, we did not provide the absolute figures but only the relations among them.

Since the analyzed journal appears in digital form and in print, there is usually a gap between the print publication date and the time when the document is put online. When not considering the one document assigned to the document type “Erratum”, FLAs also have an outstanding role here. As is exhibited in Table 3, an electronic “full length article” appeared nearly two months (50 days) before print publication on average.

Table 3. Average difference between print and online publication date (print publication years: 2002-2011) (n=321 docs)

Document type	n	Online date - print publication date (mean days)
Announcement	5	-13.2
Book review	4	-40.5
Contents list	29	12.9
Editorial Board	29	12.9
Editorial	49	9.0
Erratum	1	-145.0
Full length article	**181**	**-49.8**
Index	12	-4.9
Miscellaneous	9	32.9
Publishers note	2	-13.0
	321	-24.9

Since FLAs are the most interesting type of document from a science perspective, we performed the obsolescence analysis only for this document type. As Table 4 shows, there was a huge increase in the number of downloads between 2002 and 2011. By far the largest proportion of this increase is due to the fact that with each (download) year the range of downloadable documents increased (from 13 in 2002 to 181 in 2011). However, also the general rise in the use of e-journals between 2002 and 2011 might have partly contributed to this increase.

An analysis of the obsolescence characteristics reveals that from the downloads of a certain year, most of them allot to articles either published in the download year or one year earlier (formatted in bold). In six cases articles were already downloaded one year before print publication (in grey) since they were already available online. Accordingly, it can be concluded that more downloads accrue to recently published articles. However, also older articles are downloaded relatively often. In contrast, in our former studies in the fields of oncology (Schloegl & Gorraiz 2010) and pharmacy (Schloegl & Gorraiz 2011) half of the downloads were already made within the first two years after publication.

Table 4. Yearwise *relation*[1] of downloads per print publication year (2002-2011), (doc type: full length article, download year: <=2011) (n=181)

Pub year	n	Download year											downloads per doc – *relations*[1]
		2002	2003	2004	2005	2006	2007	2008	2009	2010	2011	All	
2002	13	1.0	**2.3**	1.7	1.3	1.2	1.4	2.4	2.8	2.8	2.7	19.6	7.4*x
2003	21	0.0	1.3	**2.2**	1.0	1.0	0.9	1.5	1.3	1.5	1.1	11.9	2.8*x
2004	17			1.7	**2.6**	2.1	2.2	2.4	2.7	2.9	2.3	18.9	5.5*x
2005	18				1.7	2.3	1.8	2.0	2.4	2.6	2.2	15.0	4.1*x
2006	14				0.2	**2.4**	2.1	1.8	2.1	2.0	2.0	12.5	4.4*x
2007	18					0.0	**2.7**	**3.6**	3.4	3.5	2.9	16.1	4.4*x
2008	16						0.0	2.9	**3.5**	3.0	2.4	11.8	3.6*x
2009	14								3.1	**4.0**	3.1	10.2	3.6*x
2010	21									3.9	4.4	8.3	2.0*x
2011	29									0.3	**5.6**	5.9	1.0*x
all	181	1.0	3.7	5.6	6.8	8.9	11.1	16.6	21.4	26.4	29.0	130.4	

[1] Since the download numbers are very sensitive, we did not provide the absolute figures but only the relations among them.

Citation data

Table 5 shows, first of all, that ScienceDirect and Scopus use different document types which are not compatible to each other. The document type "full length article" in ScienceDirect mainly corresponds to the three Scopus document types "article", "conference paper" and "review". As expected, reviews receive more citations per document (20.2) than articles (14.8) whereas conference papers received only very little citations.

One interesting fact is that more than one quarter (27%) of all documents were not cited in the citation window (2002-2011). This is mainly true for editorials (79%) and conference papers (69%). (In contrast, there allotted a certain download volume also for document types like "editorial", "book review" or "announcement" in ScienceDirect.) Also the publication date has a great influence on the citation rate. Usually only a minority of the articles are cited in the year of publication. For instance, 21 articles from 2011 did not receive any citation in the publication year.

Table 5. Distribution of Scopus document types and citations per document type (2002-2011).

Doc type	no. docs	no. uncited	% uncited	Cites	%	Cites per doc type
ar	151	22	15%	2563	86,4%	14.8
cp	13	9	69%	8	0,3%	0.4
ed	33	26	79%	13	0,4%	0.2
re	18	1	6%	383	12,9%	20.2
all	215	58	27%	2967	100%	10.9

ar=article, cp=conference paper, ed=editorial, re=review

Table 6 shows the year-wise citation distribution of articles, reviews and conference papers between 2002 and 2011. As can be seen, in all citation years – from which 2011 is the most interesting one because it has the longest time frame – most citations (formatted in bold) accrue to articles from the publication year 2002. In contrast, as was already mentioned above, only a few documents were cited in the year of publication. This shows a clear difference to downloads which have their maximum either in the year of publication or one year later.

Table 6. Year-wise citations (2002-2011) per publication year (document types: article, review, conference paper), only cited documents (n=150).

Pub year	n	Citation year											cites per doc
		2002	2003	2004	2005	2006	2007	2008	2009	2010	2011	all	
2002	13	2	19	**38**	**69**	**88**	**105**	**158**	**165**	**194**	**199**	1037	79.8
2003	14		1	6	21	27	39	35	41	40	39	249	17.8
2004	17				15	40	56	74	78	88	107	458	26.9
2005	19					16	46	78	76	93	99	408	21.5
2006	14				1	2	14	31	31	53	49	181	12.9
2007	18						1	31	74	92	85	283	15.7
2008	15							3	30	69	83	185	12.3
2009	14								3	34	57	94	6.7
2010	18									5	40	45	2.5
2011	8										14	14	1.8
all	150	2	20	44	106	173	261	410	498	668	772	2954	

Readership data

Table 7. Readership data per print publication year (2002-2011), (doc type: full length article, data extracted from Mendeley: October 2012) (n=181)

Publication year	n	Occurrences in user libraries	Occurrences per doc
2002	13	566	43.5
2003	21	344	16.4
2004	17	471	27.7
2005	18	371	20.6
2006	14	382	27.3
2007	18	580	32.2
2008	16	451	28.2
2009	14	416	29.7
2010	21	499	23.8
2011	29	537	18.5
all	181	4617	25.5

Since time stamps of the readership data were not available at the date of analysis, we could not perform an obsolescence analysis. Instead, Table 7 displays how many times (full length) articles from the publication years 2002-2011 were mentioned in total in Mendeley user libraries. Contrary to downloads and in particular to citations, the distribution of the occurrences is relatively even. One reason why older articles do not have higher readerships could be that Mendeley started in 2009 and has become popular in 2010.
Another interesting characteristic of Mendeley is its user structure. A preliminary analysis of the readers of the Journal of Strategic Information Systems revealed that by far the majority of them are students, in particular PhD students.

Comparison among downloads, citations and readership data

Figure 1 shows a medium to high relation among downloads, citations and readership data which is higher for downloads and citations (Spearman r = 0.77) and for downloads and readership data (Spearman r = 0.73). Among the ten most downloaded articles, seven (not the same) are in the top-10 readership and citation rankings. The correlation was lower between readership data and citations (Spearman r = 0.51) but in line with previous studies. For instance, Bar-Ilan (2012) calculated a correlation between Mendeley and Scopus for articles, reviews and conference papers from the Journal of the American Society of Information Science and Technology (publication years: 2001-2011) of 0.5 (data collection: April 2012). The correlation identified by Li, Thelwall and Giustini (2012) was similar between WoS citations and occurrences in Mendeley user libraries for articles having appeared 2007 in Nature (Spearman r=0.56) and Science (Spearman r=0.54) (data collection: July 2012). Only the analysis by Li and Thelwall (2012) found a higher correlation (Spearman r=0.68) between Mendeley and Scopus for 1397 genomics and genetics articles published in 2008 (data collection: January 2012). One reason for the lower correlation between Mendeley readership and citation data could be that Mendeley users have only been creating their libraries since 2009. Therefore, older articles may have lower occurrences in comparison to downloads in ScienceDirect and, in particular, to citations in Scopus, where there was the possibility to download/cite them already before 2009. Another reason could be that Mendeley users are younger (most are PhD or Master students) who prefer more up-to-date articles. This could in particular be true for computer science. One indication for both arguments could be that there was one article from the publication years 2006, 2008, 2009 and 2010 respectively in the top-10 readership ranking, while the most up-to-date article in the corresponding citation ranking was from 2005.

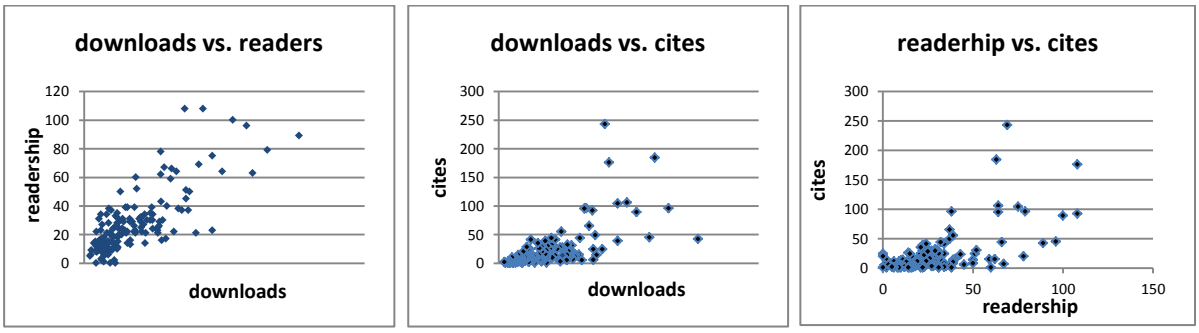

Figure 1. Downloads vs. readers vs. cites, scattergram (publication year: 2002-2011, doc type: full length article, only articles cited at least once) (n=151)

Conclusions and future research

Our analyses revealed both commonalities and differences among citations, downloads and readership data. Citations and downloads have clear differences in their obsolescence charac-

teristics. While it takes several years until articles from the analyzed journal get cited more often, the highest downloads usually happen within the first two years that follow publication. We computed a medium to high correlation among citation, download and readership frequencies. However, a rough analysis of Mendeley users suggests that its user population differs from the one having published (and cited) articles in Scopus. Since this might also be true for the ScienceDirect user community, a perfect relation among these three indicators could not be expected.

As soon as we receive time stamps for the readership data, we will start the obsolescence analyses with them. Since we are aware that the results of our study lack generality due to the small sample, we plan investigations with more journals also from other disciplines (e.g. economics, oncology, linguistics, and history) in the near future.

Acknowledgments

This report is based in part on analysis of anonymous ScienceDirect usage data and/or Scopus citation data provided by Elsevier within the framework of the Elsevier Bibliometric Research Program (EBRP). Readership data were provided by Mendeley. The authors would like to thank both Elsevier and Mendeley for their great support. The Know-Center, which is the affiliation of one co-author, is funded within the Austrian COMET program - Competence Centers for Excellent Technologies - under the auspices of the Austrian Federal Ministry of Transport, Innovation and Technology, the Austrian Federal Ministry of Economy, Family and Youth, and the State of Styria. COMET is managed by the Austrian Research Promotion Agency FFG.

References

Anonymous (2012). Alternative metrics. *Nature Materials*, 11(10), 23 October 2012, 907.

Bar-Ilan, J. (2012). JASIST@mendeley. *ACM Web Science Conference 2012 Workshop*. Retrieved January 23, 2013 from: http://altmetrics.org/altmetrics12/bar-ilan/

Bar-Ilan, J., Haustein, S., Peters, I., Priem, J., Shema, H. & Terliesner, J. (2012). Beyond citations: scholars' visibility on the social web. In *Proceedings of the 17th International Conference on Science and Technology Indicators* (pp. 98-109). Montréal: Science-Metrix and OST.

Bollen, J. & Van de Sompel, H. (2008). Usage impact factor: The effects of sample characteristics on usage-based impact metrics. *Journal of the American Society for Information Science and Technology*, 59(1), 136-149.

Hammerton, J., Granitzer, M., Harvey, D., Hristakeva, M. & Jack, K. (2012). On generating large-scale ground truth datasets for the deduplication of bibliographic records. In *International Conference on Web Intelligence, Mining and Semantics 2012* (p. 18). ACM. doi: 10.1145/2254129.2254153

Henning, V. & Reichelt, J. (2012). Mendeley - A Last.fm for research? In *IEEE Fourth International Conference on eScience* (pp. 327–328). IEEE. doi: 10.1109/eScience.

Kraker, P., Körner, C., Jack, K. & Granitzer, M. (2012). Harnessing User Library Statistics for Research Evaluation and Knowledge Domain Visualization. In *Proceedings of the 21st international conference companion on World Wide Web* (WWW 2012 – LSNA'12 Workshop) (pp. 1017-1123). ACM. doi: 10.1145/2187980.2188236.

Li, X., Thelwall, M. & Giustini, D. (2012). Validating online reference managers for scholarly impact measurement. *Scientometrics*, 91(2), 461-471.

Li, X. & Thelwall, M. (2012). F1000, Mendeley and traditional bibliometric indicators. In *Proceedings of 17th International Conference on Science and Technology Indicators* (STI 2012), (pp. 541-551). Montréal: Science-Metrix and OST.

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.

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.