Full Text downloads and citations: Some reflections

Eugene Garfield
Chairman Emeritus. Thomson Reuters Health Sciences
(formerly Institute for Scientific Information)
Philadelphia. PA.

Keynote lecture at the Seminar
"Scientific Measurement and Mapping"
Hilton Hotel , Santa Fe, New Mexico, May 10th, 2011
http://garfield.library.upenn.edu

It was with great reluctance that I agreed to be your keynote speaker. However, it was difficult to resist the entreaties of Gali Halevi and Henk Moed. And when I finally read the topics to be covered in today's program I felt somewhat embarrassed and humbled by the level of expertise assembled here today. It made me realize even more how far behind I am in the exponentially growing literature of scientometrics and mapping. I feel the same way as I contemplate the dozens of papers that I post each week on the SIGMetrics listserv of the *ASIS&T*.

I say this even though I increasingly feel that 1849 epigram of Jean-Baptiste Alphonse Karr is true: " plus ca change, plus c'est la meme chose; "The more things change, the more they are the same." What used to be readerships, now are downloads. Sixty years ago librarians were still trying to find the sources of literature use. When readers were asked to identify where they first heard about a particular paper it was usually in a list of cited references from a current paper and especially review papers. Saul Herner reported on this over fifty years ago.

That finding was one of the inspiration points in my quest for the ideal information retrieval system. Review literature continues to be a significant starting point for many.
In the time I have available, I will not dwell on the slides that Henk helped me prepare.

The key points about downloads and citations that need to be stressed are briefly:

1. The positive correlation between downloads, that is readerships or use, is partly due to the effects of citations upon downloads.

2. Initial downloads and citations hardly correlate, and relate to distinct phases in processing relevant scientific information.

3. Later downloads and citations show statistically similar properties of ageing and frequency distribution.

4. None of the studies reported in the literature were generalized across all disciplines. What may be true for information studies will not be true for biochemistry or molecular biology.

5. Steve Harnad has valiantly tried to show the connections between open access and citations. But there is not much difference between these claims and the old claims about readership versus citations.

6. Many papers are downloaded and may or may not be read and even if they are, will not necessarily be followed by a citation.

7. Unless the downloaded papers involved are related in some way to the research being reported, they will not be cited.

8). "Hot papers", that is, those that are cited early on, will have been downloaded heavily, but that applies only to a small percentage of papers, many of which go on to become *Citation Classics.*

For those of you who would like to expand their coverage of the literature on this subject, I would remind you of the annotated bibliography on "Open Access Citation Advantage" by Ben Wagner although this is not strictly speaking the same as downloads.

Let me give you an example that illustrates why things change but also remain the same.

In the latest issue of *JASIS&T* there is a paper by Frandsen and Nicolaisen from Denmark on "Testing the flattery Citations Hypothesis" in the field of economics. In it they quote a statement I made in *Current Contents* back in 1974 in which I said:

"In some cases, without doubt, citations may be used for all those sullied purposes that supposedly demean its usefulness for information retrieval: flattery, padding, borrowed distinction, etc. "However, contrary to Seglen, Garfield maintains such cases are the "the trivial exceptions." The results of our [earlier] study of four LIS journals confirm this.

Most of you have heard of the Mertonian ethos formulation known as the CUDOS norms, that is, Communality, Universalism, Disinterestedness, and organized Skepticism.

Without providing anything but anecdotal evidence there is a new generation of writers who would have us believe that the Mertonian ethos has been abandoned by today's authors. They are a throwback to the work of MacRoberts who would have us believe that citation analysis is a waste of time because authors do not adequately cite those who have influenced their work.

While I would be amongst the first to denounce those who are guilty of the so-called "Disregard Syndrome" such anecdotal and relatively rare incidents do not warrant the generalization that the literature is in decline. They should be treated with disdain along with other doomsayers who feel that the internet and Googling are spawning a new generation of scholars guilty of disregarding the earlier literature. On the contrary, there is evidence that bibliographies are longer. So much so that everyone complains of information overload. What is missing is a universal standard for traditional citation practice, but also the kind of precise citation standards that would enable us to conduct not only more precise information retrieval but also the kind of citation context analyses exemplified by the work of Henry Small and others. Perhaps he will tell us more later about sentiment analysis, but I will close by referring back to a paper I presented back in 1964 in which I asked whether citation indexing can be automated, that is, can criticism and documentation of research papers be automated. While the mechanics of assembling citation indexes had been automated with the publication the first experimental *Genetics Citation Index* and then the 1961 *Science Citation Index,* it would be another 37 years before Lawrence and Giles at NEC created *Citeseer* and called it an automated citation indexing system. It permits the

visualization of citations in context. But that is not what might be called artificially intelligent interpretation of references as is implied by sentiment analysis or even the functional analysis of citations as performed by Mike Moravcsik in 1975. While there have been great strides in the mechanical translation of languages, comparable advances in mechanical documentation and criticism have a long way to go.

In conclusion I would just like to mention the mapping work that I began with Irving Sher back in 1963. This led to our report on tracing the history of DNA up to that time. Since then I have worked with two Russian colleagues for the past ten years to develop software for creating historiographs of core citation collections. The software called *HistCite* was recently acquired by ThomsonReuters and is available free of charge from the website at www.histcite.com.

This coincides with the release of version five of the *Web of Knowledge*.

If you would like to view the numerous *HistCite* historiographs that are publicly available just go to my web site at http://garfield.library.upenn.edu/histcomp/

One particular set of graphs that may interest you are those connected with the work of Ton Van Raan http://garfield.library.upenn.edu/histcomp/index-vanRaan.html

Slides follow on page 4

Full text downloads and citations: Some reflections

Eugene Garfield

Keynote Lecture at the Seminar
'Scientific Measurement and Mapping'
Santa Fe, 10-11 May 2011

Analogy Model

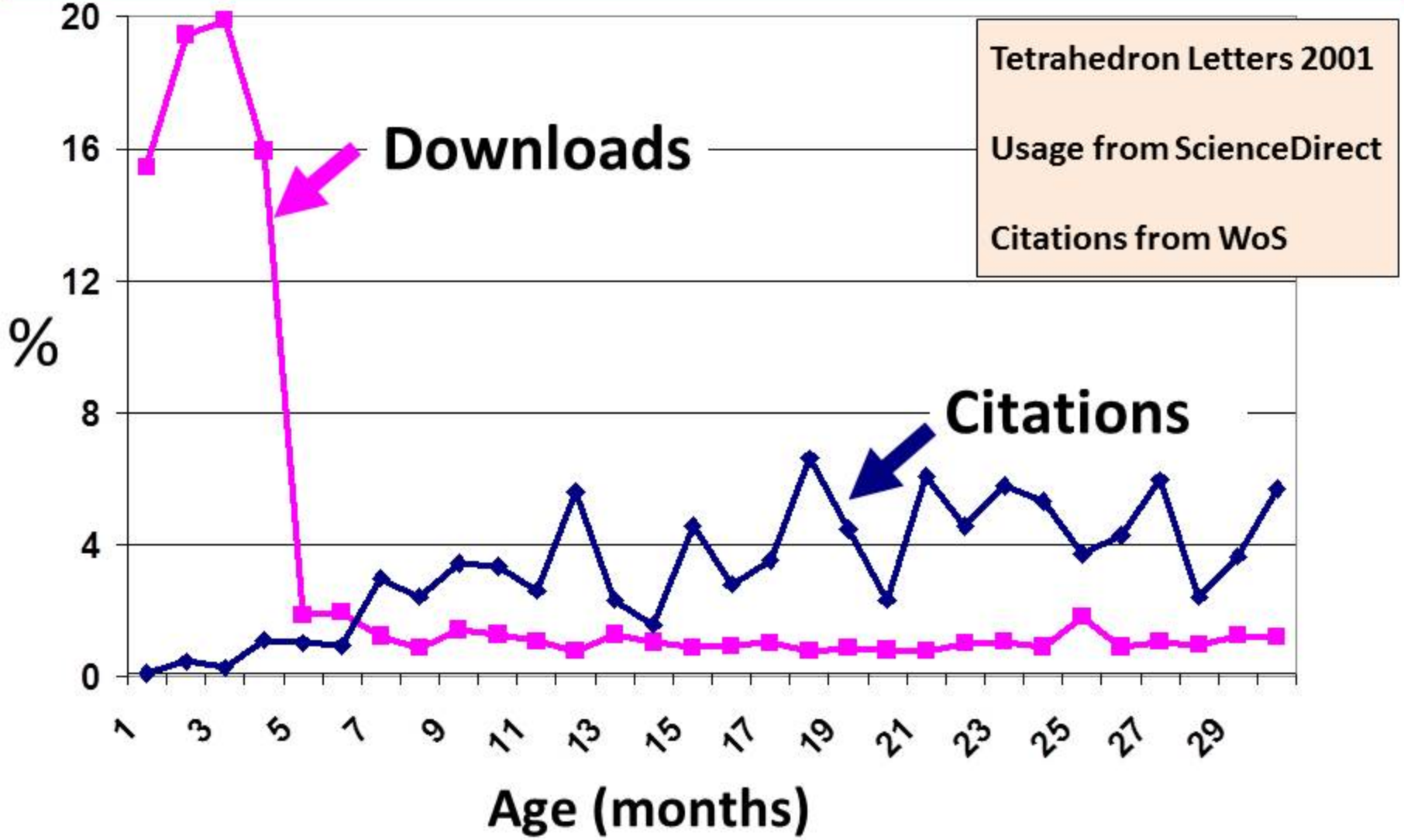
Citations (Formal use)	***Downloads (Informal use)***
Publishing authors	Article downloaders
Citing a document	Downloading the full text of a document
Article	User session
Author's institutional affiliation	User's account name
Number of times cited	Number of times down-loaded as full text

Indicators

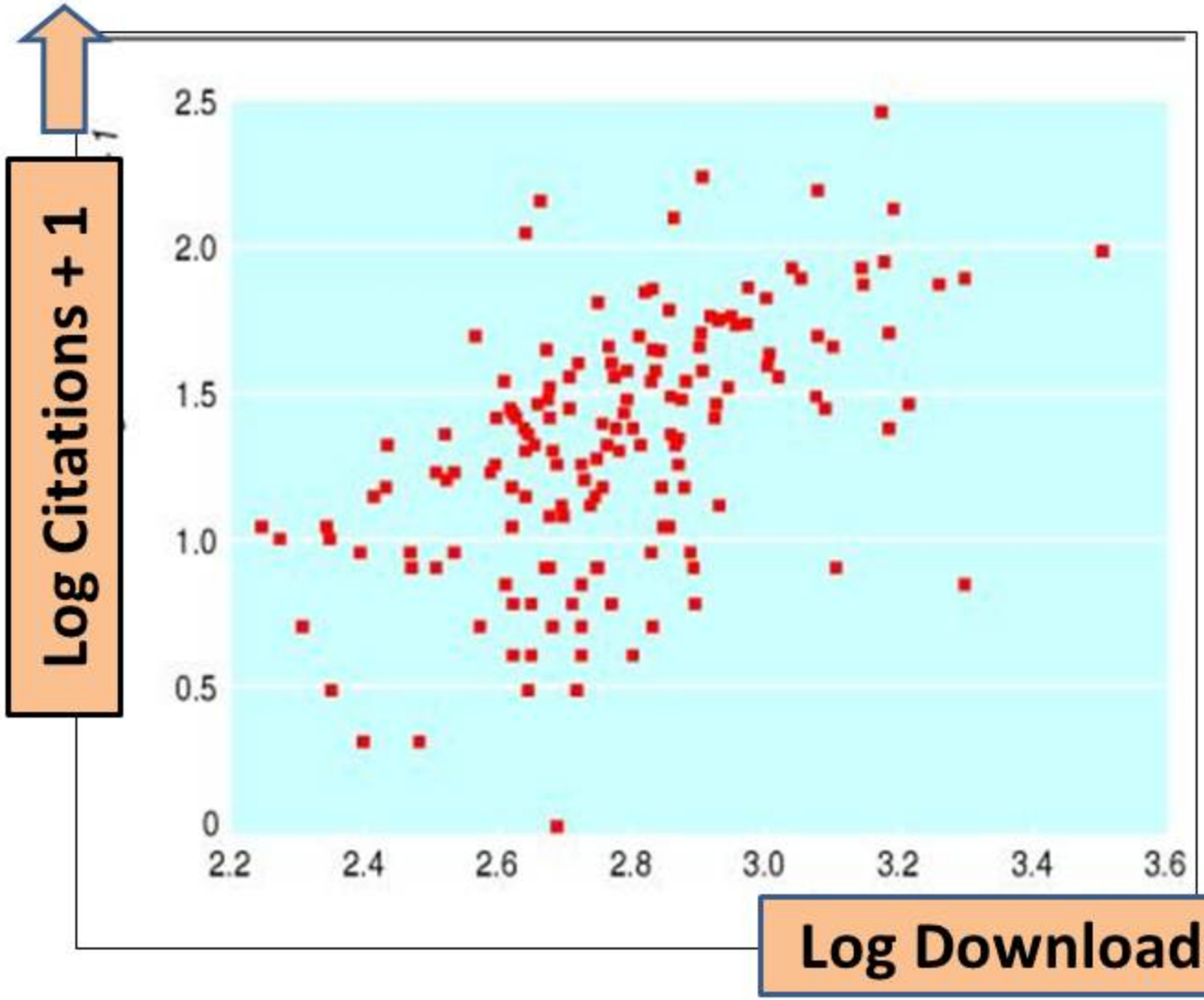
Citations (Formal use)	*Downloads (Informal use)*
Journal Citation Impact Factor (JCIF)	Journal Download Impact Factor (JDIF) [Bollen & Van de Sompel, 2008]
Journal Citation Immediacy Index (JCII)	Journal Download Immediacy Index (JDII) [Wan et al., 2010]
Journal Prestige Indicators (Bollen et al., Eigenfactor, SCImago)	?

Downloads and citations show different age distributions

[Moed, JASIST, 2005]



153 Papers in BMJ show positive correlation between downloads and citations [Perneger, BMJ. 2004]



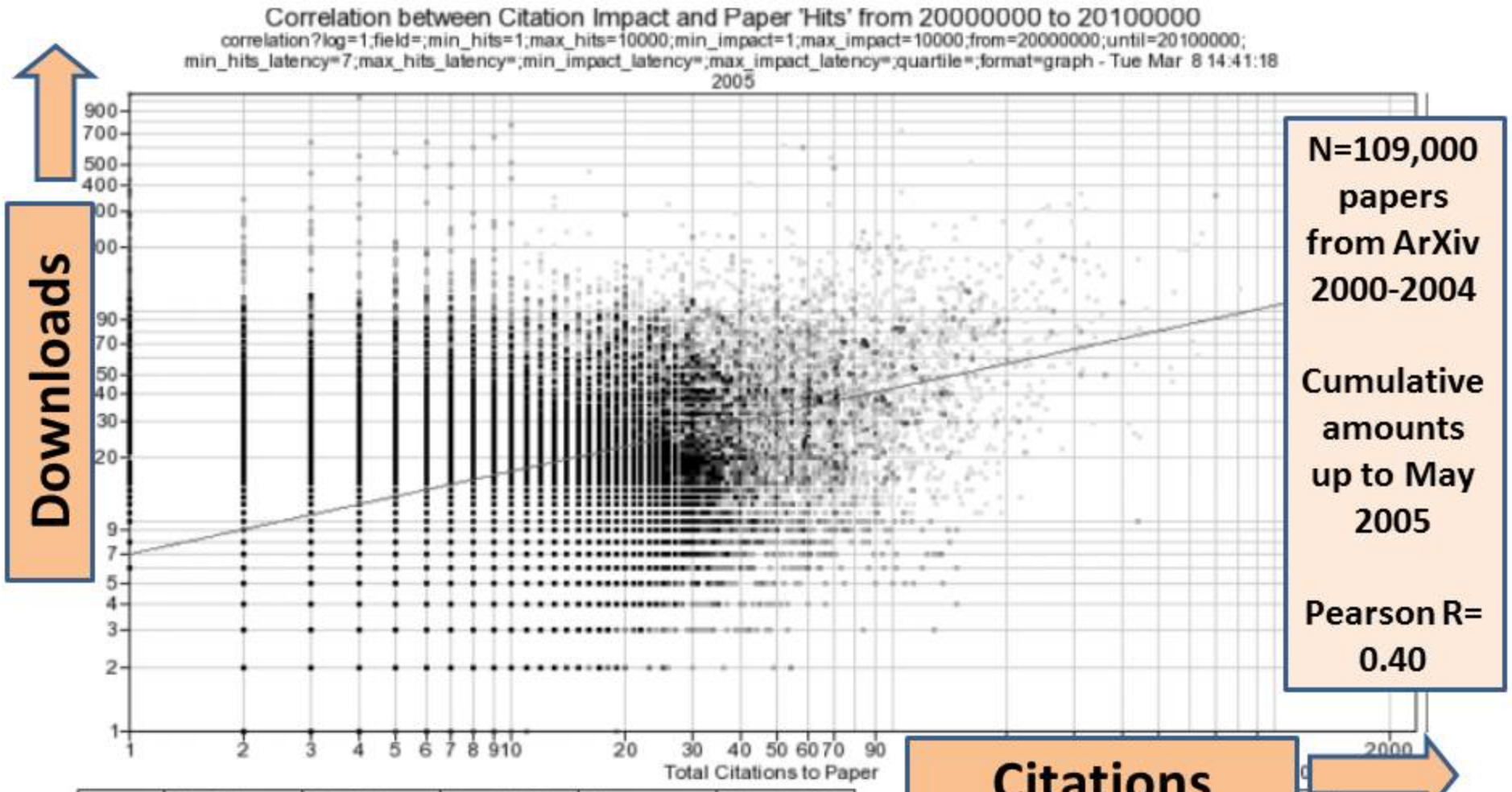
153 papers in BMJ (1999)

Downloads one week after publ.

Citations (WoS) 5 year window

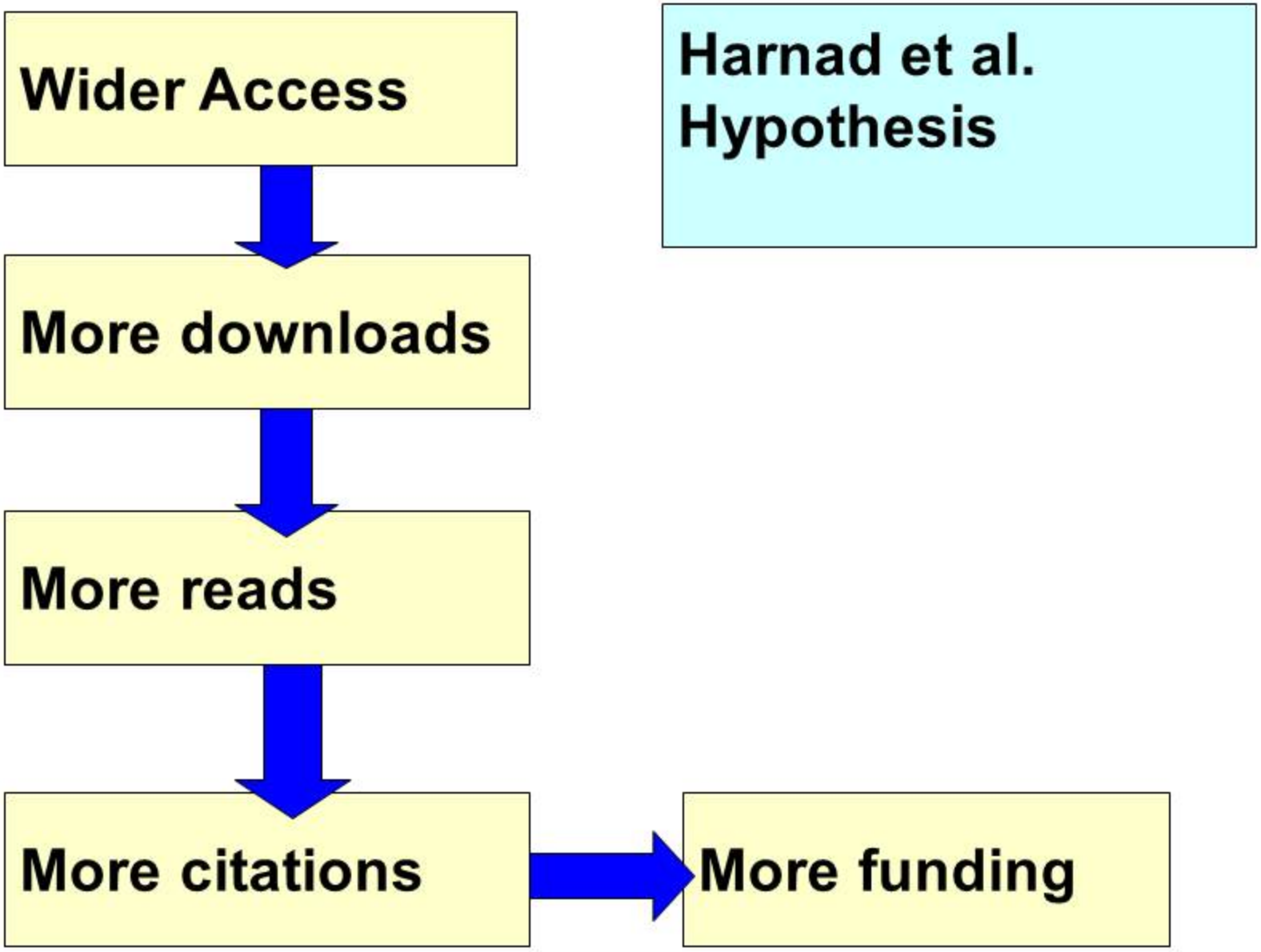
Pearson R = 0.54

Downloads and citations of papers in ArXiv correlate positively [Brody, Harnad and Carr, JASIST, 2006]



	Min Filter	Max Filter	Sum (0.53)	Mean	Standard Dev
Impact	1	10000	142442.59	1.31	1.14
Hits	1	10000	268162.56	2.46	0.76

n (t)	109101 (142.119)
Non-Dir. Sig.	<0.0001


Wider Access
More downloads
More reads
More citations
More funding
Harnad et al.
Hypothesis

Open Access Citation Advantage: An Annotated Bibliography. Ben Wagner

[Issues in Science and Technology Librarianship, Winter 2010]

- **Review articles [5 reviews]**
- **Studies showing an open access citation advantage (OACA) [39 articles]**
- **Studies showing either no OACA effect or ascribing OACA to factors unrelated to OA publication [7 articles]**

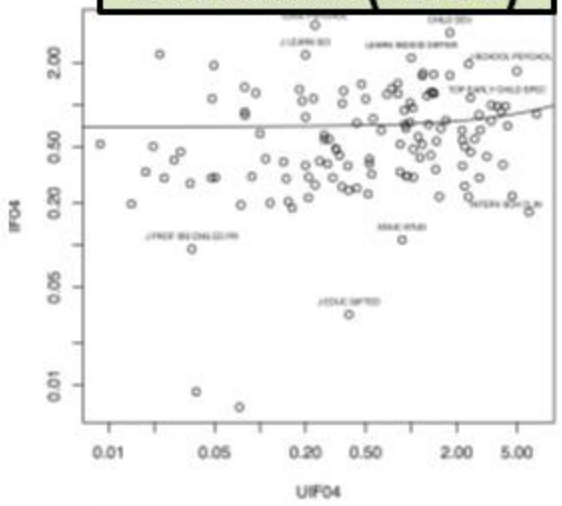
In most disciplines journals show a negative correlation between downloads citations [Bollen et al, JASIST, 2008]

Citations (JCIF)

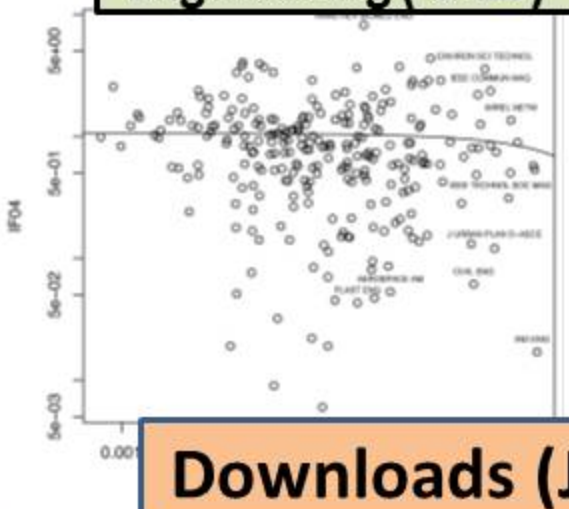
Downloads from CSU Servers

Citations from WoS

Education (+0.22)

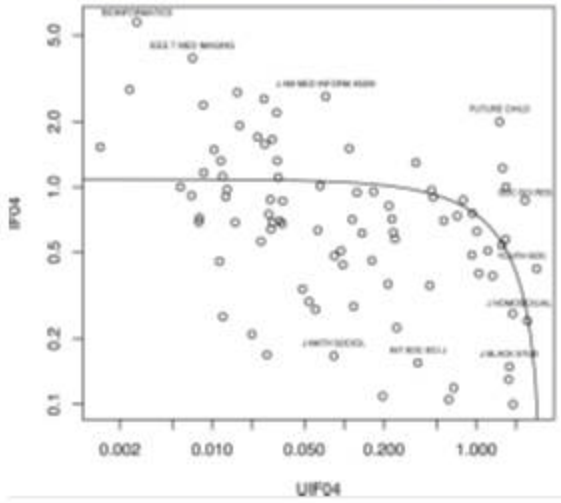


Engineering (-0.15)

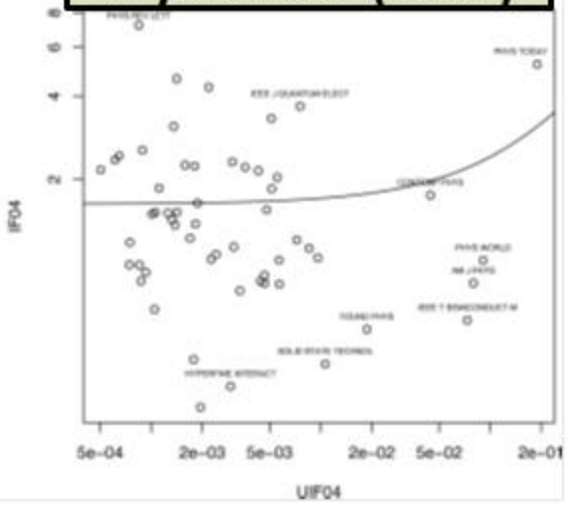


Downloads (JDIF)

Interdisc stud (-0.47)



Physical Sci (-0.22)



Conclusions: Bollen et al., JASIST, 2008

- "The particular scientific and demographic characteristics of a discipline have a strong effect on resulting usage based assessments of scholarly impact"

- "As the number of graduate students and faculty increases in a particular discipline, Usage Impact Factor rankings will converge more strongly with the ISI Impact Factor".

Spearman Rho between Download Immediacy Index (JDII) and Citation Impact Factor (JCIF) for 37 pharmacological journals [Schloegl & Gorrais, 2011]

	JCIF 2008	JCIF 2007	JCIF 2006	JCIF 2005	JICF 2004
JDII 2006	0.75	0.70			
JDII 2005		0.74	0.77		
JDII 2004			0.78	0.83	
JDII 2003				0.76	0.70

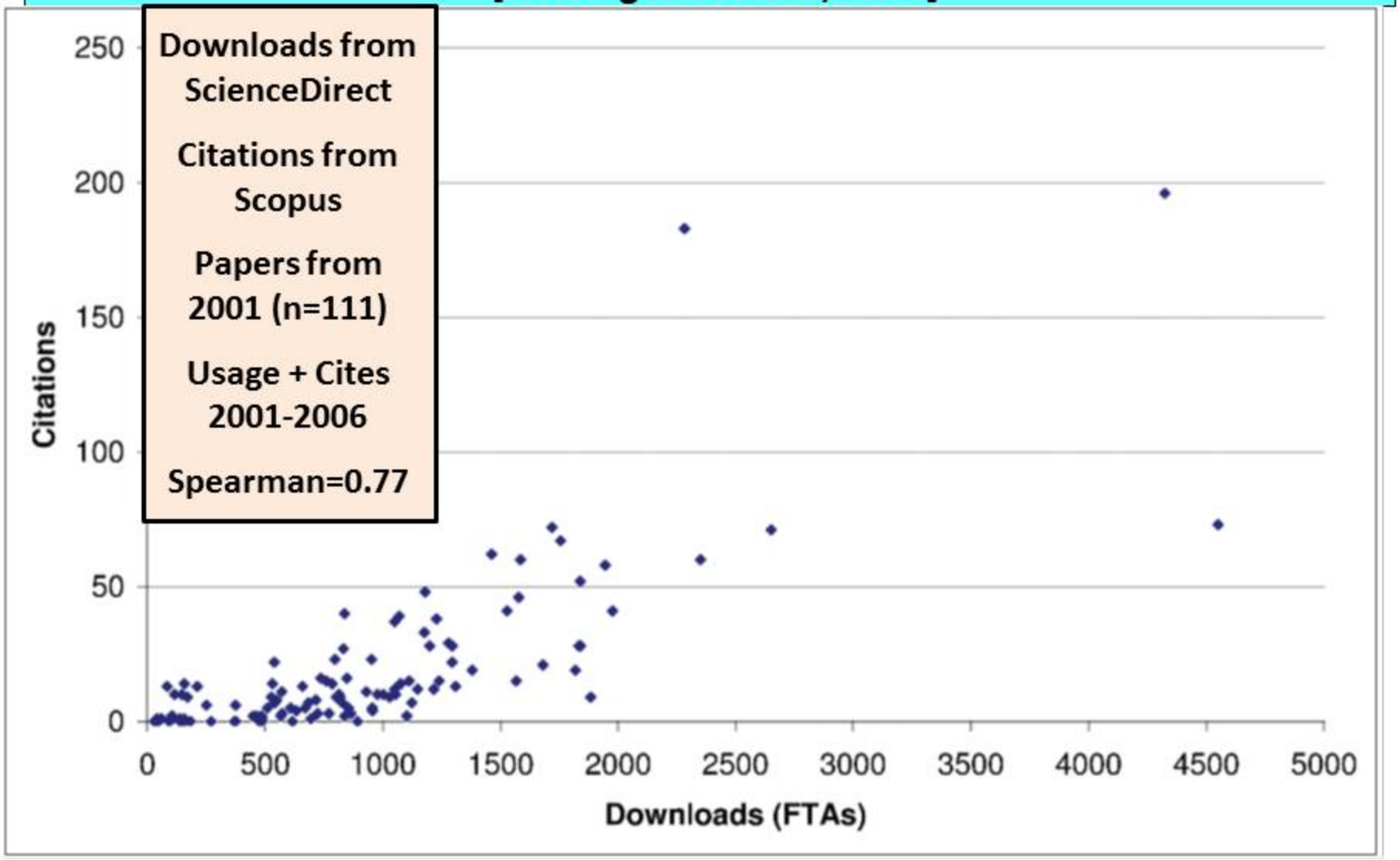
37 Pharmacol Journals both in WoS and ScienceDirect

Downloads from ScienceDirect

Citations from Scopus

Papers in Drug Discovery Today show positive correlation between downloads and citations

[Schloegl & Gorrais, 2011]



Papers in Tetrahedron Lett do not show a significant correlation between 'initial' downloads from ScienceDirect and WoS citations [Moed, JASIST, 2005]

Variables	Spearman Rho
Downloads vs. citations	0.22
'Later' (>3 months) downloads vs. citations	0.33
'Initial' (<3 months) downloads vs. citations	0.11 (n.s.)

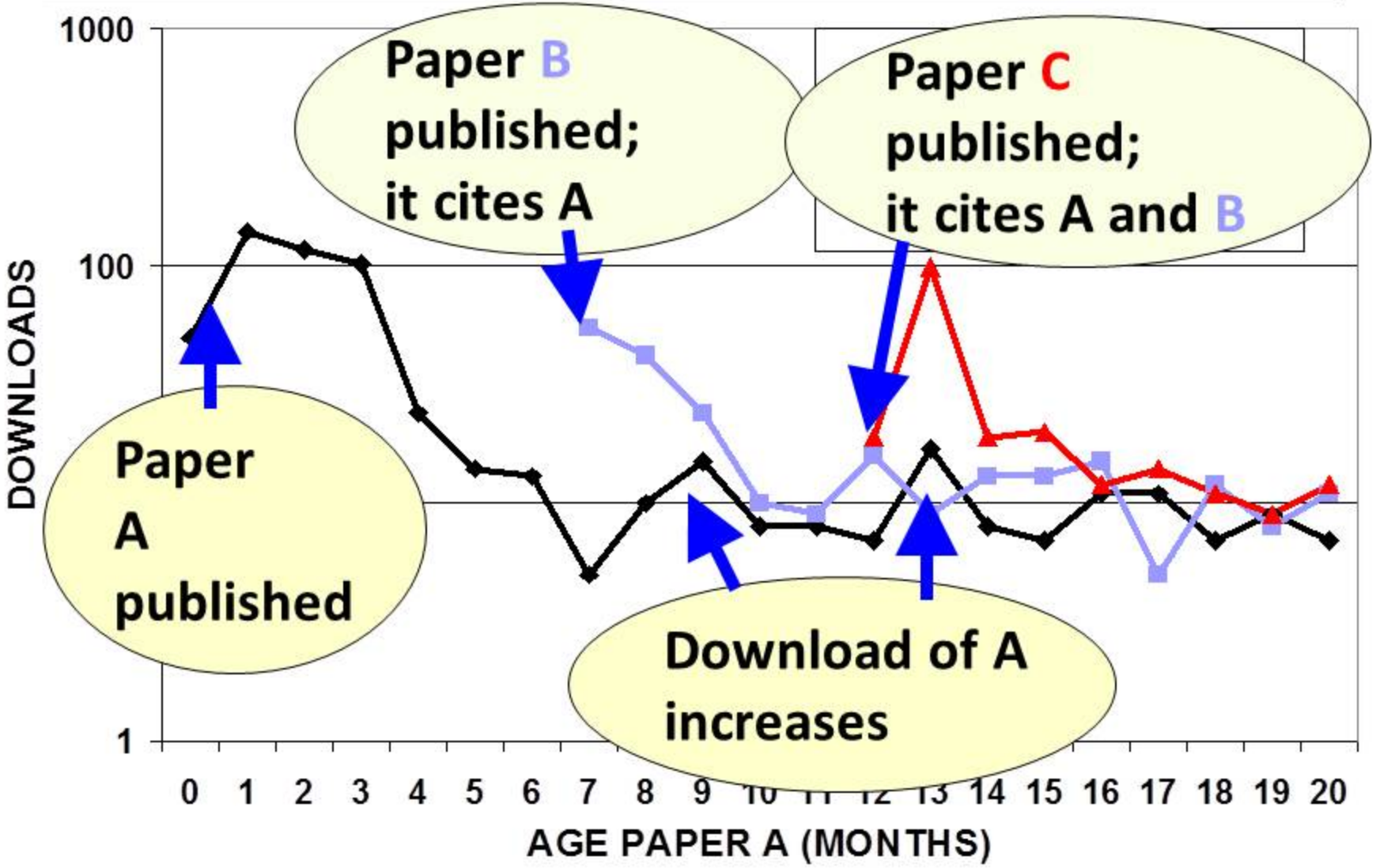
More downloads → more citations

or

More citations → more downloads?

Citations lead to downloads

[Moed, JASIST, 2005]



Conclusions [Moed, JASIST, 2005]

- Positive correlation between downloads and citations is partly due to the effect of citations upon downloads
- 'Initial' downloads and citations hardly correlate, and relate to distinct phases in processing relevant scientific information
- 'Later' downloads and citations show statistically similar properties of ageing and frequency distribution

Some reflections - 1

- None of the studies were generalized across all disciplines
- Harnad et al. tried to show the connections between open access and citations.
- There is not much difference between these claims, and the old claims about readership versus citations.
- Many papers are downloaded and may or may not be read, and even if they are will not necessarily be followed by a citation.

Some reflections - 2

- Unless the downloaded papers involved are related in some way to the research being reported, they will not be cited.
- "hot papers", those that are cited early on will have been downloaded heavily, but that applies only to a small percentage of papers.
- Relationship varies from field to field.

Literature

- 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, 136–149.
- Brody, T., Harnad, S. & Carr, L. (2006). Earlier Web Usage Statistics as Predictors of Later Citation Impact. Journal of the American Society for Information Science and Technology 57, 1060–1072.
- Kurtz, M.J.,Eichhorn, G. ,Accomazzi, A.,Grant, C., Demleitner, M. & Murray, S. (2005). Worldwide Use and Impact of the NASA Astrophysics Data System Digital Library. Journal of the American Society for Information Science and Technology 56, 36-45.
- 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, 1088-1097.
- Perneger, T.V. (2004). Relation between online “hit counts” and subsequent citations: prospective study of research papers in the BMJ. British Medical Journal 329, 546–547.
- Rowlands, I. & Nicholas, D. (2007). The missing link:journal usage metrics. Aslib Proceedings: New Information Perspectives 59, 222-228.
- 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, 161-170.
- 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, 555–566

-