Log Usage Analysis: What it Discloses about Use, Information Seeking and Trustworthiness

David Nicholas*, David Clark**,
Hamid R. Jamali***, Anthony Watkinson****

ARTICLE INFO

Article history:
Received 8 March 2014
Revised 25 May 2014
Accepted 1 June 2014

Keywords:
Log Analysis,
Usage Data,
E-journals,
Trustworthiness,
Information Seeking,
Scholars

ABSTRACT

The Trust and Authority in Scholarly Communications in the Light of the Digital Transition research project[1)] was a study which investigated the behaviours and attitudes of academic researchers as producers and consumers of scholarly information resources in respect to how they determine authority and trustworthiness. The research questions for the study arose out of CIBER's studies of the virtual scholar. This paper focuses on elements of this study, mainly an analysis of a scholarly publisher's usage logs, which was undertaken at the start of the project in order to build an evidence base, which would help calibrate the main methodological tools used by the project: interviews and questionnaire. The specific purpose of the log study was to identify and assess the digital usage behaviours that potentially raise trustworthiness and authority questions. Results from the self-report part of the study were additionally used to explain the logs. The main findings were that: 1) logs provide a good indicator of use and information seeking behaviour, albeit in respect to just a part of the information seeking journey; 2) the 'lite' form of information seeking behaviour observed in the logs is a sign of users trying to make their mind up in the face of a tsunami of information as to what is relevant and to be trusted; 3) Google and Google Scholar are the discovery platforms of choice for academic researchers, which partly points to the fact that they are influenced in what they use and read by ease of access; 4) usage is not a suitable proxy for quality. The paper also provides contextual data from CIBER's previous studies.

1. Introduction

In a complex, borderless, multi-platform scholarly environment metrics and algorithms are increasingly being utilised to deal with the information tsunami, to sort the wheat from the chaff, the

* CIBER Research Ltd, UK (dave.nicholas@ciber-research.eu)
** CIBER Research Ltd, UK (david.clark@ciber-research.eu)
*** CIBER Research Ltd, UK (office@ciber-research.eu)
**** CIBER Research Ltd, UK (anthony.watkinson@btinternet.com)
International Journal of Knowledge Content Development & Technology, 4(1): 23-37, 2014.
http://dx.doi.org/10.5865/IJKCT.2014.4.1.023

1) http://ciber-research.eu/CIBER_projects.html

relevant from the irrelevant and to establish trustworthiness. Traditionally metrics have been compiled from citations (e.g. impact factor, h-index), but there are increasing moves and pressures to embrace usage (COUNTER, 2014), and, as a nod in the direction of the increasing importance of the social media, likes, comments, recommendations and bookmarks (PLOS, 2014). This paper focuses on the value of usage data as provided in access logs and, together with interview and questionnaire data, unpacks it in order to see how it stands-up as an indicator of information seeking behaviour and what can be read into it in terms of trustworthiness. The research upon which the paper is based comes from a major, international study on trust and authority in scholarly communications in science and the social sciences in the digital era. The research was funded by the Alfred P. Sloan Foundation and took place between 2012 and 2013. This research has been supplemented by data obtained by a number of studies of scholarly publishers' platforms that CIBER has been engaged in over the past five years.

CIBER's body of work on deep log research into the usage behaviour of scholars in the digital environment informed the Sloan research project and helped set its research objectives and questions (Nicholas, 2010; Nicholas et al., 2008; Nicholas and Rowlands, 2008). The work showed that the digital transition had led to fundamental changes in the way scholars, and especially the young ones, sought and selected information; that the future researcher, those born digital, might bring with them new values, perceptions and behaviours which would challenge the scholarly establishment and its existing practices, especially in regard to determining what is trusted and trustworthy.

So five years on since the original CIBER research the Sloan project provided a timely opportunity to look for the winds and seeds of change in the scholarly communications field in regard to matters of determining trust and authority as a consequence of the digital transition. The broad aim of the research was to examine how emerging digital behaviours are challenging and, perhaps, changing long-held concepts of trust and authority. The study examined how researchers assign and calibrate authority and trustworthiness to the sources and channels they choose to use, cite and publish in.

As previously mentioned, this paper focuses mostly on one aspect of the study, a log analysis, which was undertaken at the start of the project in order to build an evidence base, which would help calibrate the main methodological tools used by the project - interviews and questionnaire. Information on the whole study can be found on the CIBER website (CIBER and University of Tennessee, 2014).

2. Aims and objectives

The main purpose of the study was to identify and assess the digital usage behaviours ('footprints') of scholars, especially those that tell us about decision making and raise trustworthiness and authority issues. Of course, the very act of using something can say something about its trustworthiness and authority and this assertion is the subject of examination in an information-rich digital environment. In particular the focus is on two specific usage behaviours that would logically appear to provide data on trustworthiness and authority and hence worthy of investigation in some depth, they are: a) levels of user engagement, and, b) the use of search engines. The former discloses the extent

to which sources are heavily or regularly used, or not as the case might be, something which would point to the trustworthiness of sources made available and their 'stickiness'. The latter provides evidence on whether ease and speed of access was influencing scholarly information seeking and reading behaviour; in other words, whether scholars were choosing sources as much for their digital visibility and prominence (convenience factors) in hit lists as their authority and also whether generic search engines were usurping the products of traditional publishers and libraries and becoming the new trusted scholarly discovery brands.

The research reported here and its research questions arise from log data studies conducted by CIBER on the virtual scholar (CIBER, 2014a; Research Information Network, 2009; Nicholas et al., 2008). These studies have raised profound questions about scholarly Web behaviour and occasioned observers, such as Nicholas Carr (2011), to claim that the Internet might be making us stupid. Broadly speaking such claims arise from log findings that show tremendous volumes of digital 'activity' but also show that much of this activity appears to be 'lite' in nature; behaviour tends to be characteristically bouncing and promiscuous; and 'skittering' is the word that best describes it. Overall visits are many, but individual repeat visits are few. They tend to be short (possibly sweet) and the number of pages viewed very low, in the ones, twos and threes normally. 'One shots' - one page viewed and one visit made, tend to predominate. Search engine searching is also dominant and this explains much of this behaviour. It does not appear from the literature (see section 4) that anyone has examined logs from a trust and authority perspective, never mind evaluated whether dumbing down claims have any foundation whatsoever.

3. Methods

The raw log data for the digital library of a medium sized, international scholarly publisher that wished to remain anonymous because of commercial sensitivities, was supplied by a third-party Internet host for the year 2013. This consisted of a raw http log recording traffic. The data were parsed to remove redundant fields and requests generated by web-crawlers and other robots. Some analysis was performed as a by-product of this text parsing process. The consolidated log was then normalised as an SQL database. Subsequent analysis was performed by a combination of SQL queries, Python programs, and various spreadsheet and statistical applications.

Log data are not easily transformed into users and usage. Thus user identification, usually through IP addresses is problematical because of multi-user machines, dynamic numbering and robots. Use, typically page views, the main usage currency, is somewhat less problematic although a lot of sifting and sieving has to be conducted in order to come up with actual user content page views. More details of the method and its attraction can be found in Nicholas and Clark (2012) and Nicholas, Clark and Jamali (2014). The dataset, covering a full month's web access to more than 650 scholarly journals and over 30 million authentic page-views, was large enough to present some challenges in processing and to provide a large enough evidence base.

In the way that the log study informed the questions used for focus groups, critical incident interviews and the questionnaire in the wider study, in turn these instruments also yielded data

that provided a context and understanding of the log data. This triangulation of data adds considerable robustness to the study.

The self-report data, which is selectively used in this paper to triangulate and explain the log data, comes from:

a. Fourteen focus groups, which were held during 2012/2013 in the UK (8) and US (6). A total of 66 academic researchers, 36 from the UK and 30 from the US, attended. The purpose of the focus groups was to scope and define the second and third data collection stages of the project: critical incident interviews and questionnaire.
b. One-to-one critical incident interviews conducted with US (42) and UK (45) academic researchers, which drilled deeper into matters raised in the focus groups by asking academics about a recent paper they wrote.
c. A global questionnaire, which sought to take the opinions and data raised in the qualitative work to a much larger and international research population. More than 3500 people responded.

(More details of these methods can be found in the project report CIBER and University of Tennessee, 2014b).

A cautionary note should be made in regard to the interpretation of the data. The log data covers all types of scholars - researchers, teachers and students, whereas the self-report data pertains mainly to academic researchers, including doctoral students. The fact that researchers have been found to be the large majority users of the journal literature means that this minimises the problems this raises (Research Information Network, 2009).

4. Literature review

We know from past studies, mainly qualitative studies and questionnaire surveys, that after finding an article topically relevant, scholars look at other characteristics to check the trustworthiness of an article and decide whether to read it and then cite it or not. For example, Tenopir et al's (2011) survey showed that scholars look at prestige of the journal; the reputation of the authors; the institutional affiliation of the authors; the type of publisher of the article; the online accessibility of the article; and the source of the article (refereed journals or non-journal sources). These are, however, the actions scholars say they do or they think they do. There is little actual evidence presented in the literature as to whether scholars actually do take these actions or whether they might for instance compromise the quality and trustworthiness of the source for its ease of access. The lack of evidence might be due to the difficulty of conducting such studies and the general dearth of log studies.

Previous log studies show that Google is an extremely popular means of accessing journal content (Nicholas, Clark, Rowlands, & Jamali, 2009; Jamali & Asadi, 2010). As a matter of fact, a high proportion of researchers accessing journal articles come to a journal/publisher website via a third-party site, and this is especially so in the case of the Life Sciences (Nicholas, Rowlands, Huntington, Jamali, & Salazar, 2010). This fact might be a reason for the popularity of abstracts since users

navigate towards content in cyberspace through search engines and gateways (Nicholas, Huntington, & Jamali, 2007), of course popularity of abstracts is also partly because they provide a quick and effective means of assessing relevance of content. Past studies also revealed that those employing an alphabetic browsing list to navigate towards content were the most likely to view only a full-text article, while those using search engines were more likely to view abstracts, maybe because search engine users has a greater number of pages to view and they resort to the abstracts to make a quick decision (Nicholas et at., 2008). Another reason might be that browsers are more likely to be those with full-text access, while searchers might be from anywhere and not only those subscribed to journals. As Howard (2012) reported JSTOR registers 150 million failed attempts every year to gain access to articles they keep behind the paywall.

As Town pointed out, when people talk about online usage based on logs, it is generally access not use that they are talking about (Town, 2004). The online information-seeking process dictates that you print or download first and then take the decision about relevance later. Therefore, inevitably, a good number of full-text downloads will never be printed out or read. This of course is rapidly changing as a result of the growth in smartphone Internet access. Also some groups of users are more likely to access full-text, this is particularly true of students, who would not have had as good access to scholarly communications as members of staff and would have to face print charges if they wanted to read offline (Nicholas et al., 2008). Overall, then a review of the literature shows that not much effort has been expended in studying the trust element of scholarly communication using an evidence base.

5. Results

The analysis of the publisher's usage logs for 2013 furnishes the digital 'footprint' evidence for hundreds of thousands of scholars. The significance of these footprints will be examined from a trust and authority perspective, especially in respect to two key behaviours, levels of engagement (activity) and search engine use.

5.1. Levels of engagement

Six metrics provide a comprehensive picture of activity levels: page views, download, visits, duration of views/visits and page views per visit. Together they point to a scholarly communications environment that is very active indeed; an environment where clearly the peer reviewed journal is hugely popular information source, but also an environment where much of the activity is navigational and light.

- **Page views**.

  The publisher's digital 'library' of over 650 journals attracts well over 1 million page views per day. The publisher's peer reviewed articles are indeed immensely popular and sought. High levels of activity indeed, but there is a huge variability within that. Ten journals (1.5% of all the journals)

accounted for more than 12.6% of page views and just 25 (3.8%) accounted for 21% of all views. So even in a multi-subject environment there is a good deal of concentration in use. The most popular journal attracts more than 25,000 page-views per day and the least popular obtains fewer than 30 page views per day (Table 1). The important question of course is whether any of the variation is down to quality; that is whether you can equate popularity with quality? A simple, relatively crude test to gather evidence was undertaken. The top ten most viewed journals and the bottom ten least viewed journals were compared in respect to their impact factors. The result: while there was a slight tendency for journals with a high impact factor to be in the top most used group, there were some journals in the lowest group which had high impact factors also. Also the ones in the lowest group tended to be specialised with small audiences or newly published and this might well have determined their lowly position. So there was no clear-cut evidence showing that highly-rated journals, as defined by citations, are more popular.

- **Full text downloads**.

  Downloads are considered to be as close you can get to a satisfaction metric in the logs, and that makes it a possible indicator of trustworthiness. Around 35,000 full text articles are downloaded per day on the publisher's website; however, just three percent of visits result in a download, showing how very discriminating scholars really are when making their information choices (Table 2). Downloads though are dwarfed by abstract views; nearly 666,000 a day. This is a consequence of their navigational and fast information properties, and the fact they are free to view for all and a possible surrogate for the full-text. From the logs then it appears that scholars often check and cross-check to establish relevance and trustworthiness, and abstracts greatly helped them in the task. Ironically perhaps, in a full-text-rich environment abstracts have never been so popular.

- **Visits.**

  The website receives about 150,000,000 visits per year, another sign of very high levels of scholarly activity and the popularity of the journal as a source of scholarly information.

- **Duration.**

  The average visit lasts less than two minutes (106 seconds), which confirms the fast and brief searching style of most scholars on scholarly databases, which has been commented upon elsewhere (Nicholas et al., 2010).

- **Pages viewed per visit.**

  The average visit sees less than four pages viewed. However, the averages conceal considerable variation: 37% of all visits are to a single abstract page, another 15% bounce from the home page, another 15% bounce on the various RSS feeds.

Table 1. Page views by journal title. Top and bottom titles (one week, March, 2013)

Top 10 Journals	n	%	Cum %
1. a medical journal	25130	2.3	2.3
2. a psychology journal	19297	1.7	4.0
3. a psychology journal	18585	1.7	5.7
4. a medical journal	14388	1.3	7.0
5. a management journal	12787	1.1	8.1
6. a psychology journal	12712	1.1	9.2
7. a biology journal	10992	1.0	10.2
8. an educational journal	10901	1.0	11.2
9. a social science journal	8236	0.7	11.9
10. an applied science journal	8101	0.7	12.6
Bottom 10 journals			
1. an engineering journal	25	0.003	0.003
2. a medical journal	26	0.003	0.006
3. an engineering journal	28	0.003	0.009
4. an educational journal	30	0.003	0.012
5. an engineering journal	62	0.007	0.019
6. a medical journal	67	0.007	0.026
7. a sociology journal	68	0.007	0.033
8. a medical journal	74	0.008	0.045
9. a humanities journal	82	0.009	0.054
10. a humanities journal	89	0.009	0.063

Table 2. Type of page viewed (one week, March, 2013)

Page viewed	Number of page views	% of views
ABSTRACT	665908	58%
HOME	287356	25%
RSS feed	76442	7%
TOC ISSUE	46563	4%
FULL TEXT	35005	3%
ALERTS	7398	1%
OTHERS (14)	24291	2%
TOTAL	1142963	100%

Fully three-quarters of scholarly visitors leave this very light digital bouncer's footprint. This provides further confirmation of the dominance of the bouncer as a web user (Nicholas et al., 2004). This is not to say that there is no signs of more engaged searching being conducted, but there is not a lot of it. There is a core of heavy users, albeit small in number. Thus less than a quarter of all visits are multi-page in type, the average number of pages viewed being six, average visit time seven-and-a-half minutes. Those people who start browsing at the Table of Contents

of an issue show quite high levels of engagement: an average of 22 pages per visit. However, the numbers of these visits are quite small (600 per day).

There is another metric, which arguably ought to offer a seventh indicator of engagement and that is return visitors. It is however a very problematical metric. The returnee problem is fundamentally one of identification: if we cannot easily identify a user in the first place how can we possibly recognise one who returns. The problem is inherent in the nature of the world-wide-web. Technically, it is built on a protocol —HTTP— that is 'stateless'; each message is a one-off, not a chain of transactions. Socially, an origin in the free exchange of academic information has created an expectation that resists incentives to registration and payment. The technical solution has been 'cookies': essentially storing an identifier that accompanies each message thus linking the chain. The behavioural solution lies in various incentives to sign-up, provide personal data, and stay logged-on. There are though, three significant problems associated with cookies: 1) cookies can be cleared out between visits; 2) without a sign-on, identification is by device rather than user; 3) visitors may resist registration or provide misleading data if forced to register. All this generally means that logs, like the publisher's ones referred to in this study, provide little in the way of robust data on returnees.

The commercial viability of services such as Facebook etc. rests on the relative success they have had in overcoming this resistance to providing logged-on identification. The same may be said for various 'apps'; wrapping access, information and presentation into a single package rather than using an independent browser to present data. In both cases the effect can be seen to undercut the free and open foundation of the web: user data is gained but it is not free.

What then does this online behaviour described above tell us about the trustworthiness and quality of digital scholarly content today, especially journal articles which the publisher's database mainly covered? It would be easy to jump to the conclusion, because of the prevalent 'bouncing' behaviour, that a good number of users do not like what they are finding, perhaps because of poor or mediocre quality or simply the fact that much of it is irrelevant. If that was indeed the case, in regard to the former, this could be a consequence of the huge reach of the web and the great expansion in scholarly publishing, which has brought about a lowering in overall quality, by bringing in many more people whose contributions are not very significant. In regard to the latter this could be a consequence of the shotgun approach of search engines, which generates much noise (more on this later). There are, however, other possible explanations for this form of behaviour, some of which point to a somewhat less downbeat picture:

- Scholars just do not have the time to wade through the information flood; they 'smash and grab' as a result, which says something about the speed and hurriedness of the evaluation process.
- It is a reflection of massive and changing choice, creating an inevitable churn in use.
- It could be a function of easy and always-on access, which engenders a kind of topping up or snacking form of information feeding.
- It is a direct result of end-user checking, because in a disintermediated environment, remote from the library, users have to make trust and authority decisions themselves, and they do this by sampling the abstracts and hit lists. We are all librarians now.

The interviews and focus groups provided strong support for the views furnished above. Certainly when researchers begin their research/develop their research questions they examine lots of sources—they bounce around such as you see in the logs. As they develop their research query and obtain a firmer idea of what they are looking for their searching behavior becomes more focused. Many also mentioned the overwhelming amount of sources on the Internet/electronic databases etc. and that, too, could account for their sporadic behaviour. They follow citation hyperlinks/twitter links etc. down the "rabbit hole" just to make sure they are not missing something. After all it is much easier to open an article just to check it out (even if you have no intention of really reading it or using it) in the digital age then it was when you actually had to walk to the library and open a book/journal.

There were differences between groups of researchers. There was certainly evidence that early career researchers snack on Google, twitter and Facebook in order to assemble the material they need to look at and organise. On the other hand established researcher used a small number of sources very effectively and in detail for specific projects. They may have conducted a bit of 'topping up' at the end of projects just to add a few new touches to the paper.

Of course, there is also the possibility that the logs do not tell the true or full story about the user's journey? For instance:

- The 'session' (aka 'visit'), a key behavioural metric described above, does not mean so much as it once did now that we have the Web and always-on broadband: there is no certainty that between the first and last recorded request from a user that any consumption is happening. Nor, for that matter, what appears to be a series of entirely separate transactions might be from the end user point of view part of a single 'session'.
- Users might have downloaded and save consumption for latter or just alighted upon the exact paragraph they were looking for: we just cannot tell if a bouncer is just flitting through or has mined a nugget of deep knowledge on that first pick.
- Users are highly likely to be multitasking, conducting tabbed browsing. Thus if we see a series of rapid requests for pages, each a single step from a portal or search result, it might be evidence of rapid, but profitable, skimming on some scale, or it could be a version of saving for later: opening up each possible relevant result in a separate tab and then reading each at leisure.
- We have the problem that today's heavily scripted web-pages routinely pre-fetch data: there is no longer certainty that a continual stream of data requests can be taken as evidence of active user engagement. Some of these scripting techniques can be used to log extra information, but by adding to the stream involuntary activity that surrounds each 'real-user' action this extra information is also adding to the noise.

The log record then is linear, but we cannot assume that the activity of the user, nor the 'consumption' of content, followed that line. The sequence of content requests is not necessarily the reading order. At the individual (forensic) level we can possibly reconstruct the probable interaction. But if we attempt that on a more generic and aggregate basis we find there is too much data. The possible patterns multiply and it is difficult to decide how to aggregate them into a manageable set of

characteristic personas.

What then did focus group participants and interviewees think of usage as a metric that might help decision-making on what scholarly content to use? Well, it is clear that many researchers were unaware of all the possibilities on offer, so they were not talking on the basis of any real knowledge or experience. Those who ventured an opinion were largely negative: 1) usage counts were thought to be too easily gamed; 2) highly used articles were not said to be the best ones (when compared to editorial opinion); 3) downloads did not represent readings, because many were not read once they were downloaded; 4) usage was not a measure of good science or research but rather a consumption or popularity indicator, and media exposure could easily and massively raise an article's count.

Questionnaire respondents were of a similar mind. There was a general agreement that usage (and social media) derived metrics were indicators of popularity and not quality or credibility, and, as such, of no real help to researchers. Older researchers were more likely to believe this. There was also a significant difference in response according to the roles that researchers had. Those who worked as editors, on editorial boards or as referees (agents of the system if you like), who would be older researchers anyway, felt more strongly that usage and social mentions were mostly indicators of popularity. Researchers from less developed countries were less negative about usage metrics. Interestingly, and inexplicably, male researchers were also more likely to view usage metrics as simply popularity counts.

5.2. Search engines and discovery

What then of the reported widespread prevalence of search engine searching by scholars (Jamali and Asadi, 2010)? The referrer logs provide very specific details of search engine searching and the logs show that twenty two percent of traffic (32 million visits per year) emanates from just one search engine, Google; another 13% (around 20m visits per year) from Google Scholar. Clearly Google is a popular and, seemingly, trusted search pathway. Even the third most popular referrer is effectively a search engine, The National Center for Biotechnology Information gateway, which is responsible for about one third of a million visits. By way of contrast, despite the widespread popularity of the social media, Wikipedia and Facebook (all with hundreds of millions of users) are relatively well down, each accounting for over a quarter-of-a-million visitor referrals, suggesting possibly relatively low levels of scholarly trust in these channels or maybe just the start of something bigger. Interestingly, it would seem that people coming into a scholarly publisher's website from an academic institution website, once the normal route, is not as common as one might have expected. Thus each year there are just 0.5 million visits referred from web-pages within the .ac.uk second-level domain and 1.5m from the .edu domain. It would appear then that most scholars are not viewing the publisher's digital 'library' fully, they are in fact seeing bits of it ('snips') through some form of portal, which may be a federated search, a subject gateway, but most likely Google. Past studies also showed that third-part sites play a significant role in directing users to journal websites (Nicholas, Rowlands, Huntington, Jamali, & Salazar, 2010). This raises questions about the role of the library in the search and discovery process and, especially in the case of early career researchers who

are quite unsure about whose information they really are using (and trusting).

There is a relationship between low levels of site engagement and search engine use because search engine users are typically one-shots or bouncers (one page, one visit). Around three quarters of visits are of this kind.

The heavy use of Google and Google Scholar seen in the logs was confirmed by the interviews and questionnaire. Thus Google Scholar, proved to be the most popular discovery system, regardless of discipline, and, importantly, is regarded as a trustworthy source. Apparently more so than library websites; federated search engines and publisher platforms are rarely mentioned. This can largely be put down to the fact that the 'trusted big fat information pipe' that researchers are connected to for usage purposes is the internet and Google indexes the internet. Google Scholar and the Google search engine provide ease of access (one stop) and very, very wide information horizons. Researchers said they were influenced in what they used or read by ease of access and young researchers are most influenced. This obviously helps explain the widespread popularity of Google Scholar and the Google search engine with researchers. It is absolutely clear from all this that researchers do not live in the proverbial ivory tower. Their modes of searching are much the same as the modes of searching that we find among members of the public.

The value of search engines was also mentioned by interviewees for finding supporting references for the papers they were writing; so they have a citation, as well as usage, value. They also like the 'borderless' aspect of the search engine. Thus many researchers start with Google or Google Scholar then switch to a more specialized database, such as PubMed Central, when they had a more defined search query. In some cases, researchers found appropriate databases in an unfamiliar field, starting with Google. Researchers also use Google to check the credibility of an author.

When questioned in focus groups and interviews, researchers are generally reluctant to acknowledge that in a world of plenty, always available information, there might be the temptation to grab the first document they see. No doubt part of the reason for this arises from the fact that they would not want to readily admit to this temptation. However, a number of social scientists did air their concerns, saying that the ease of searching carried with it a risk that they might be sacrificing quality for speedy discovery. They admitted to using what they could get hold of most easily: for instance, what the library has access to, or, they could obtain directly from the author via an email contact they had. In the words of one social science researcher, 'Any barrier will put you off chasing something'. While few researchers admitted to choosing research material on the basis of its prominence on a search engine hit-list, they readily admitted to doing this in regard to non-research activities, especially university administrative and management activities; and leisure activities could be added to this list.

Researchers, in the more anonymous environment of the questionnaire, were more likely to say they were influenced in what they used or read by ease of access factors, but not heavily so. Thus, the levels of agreement to both of the following statements were around the 30% mark: a) If the information is not central to my research area, the ease of availability of a source is more important than its quality; b) When pressed for time, the ease of availability of a source overtakes considerations about its quality. What was very interesting from a Google Generation perspective is that younger researchers were more likely to say that ease of access was a factor

in what they used. Not surprisingly then younger researchers were the biggest users of search engines.

6. Conclusions

By relating evidence based log usage data with self-report data it has been possible to show that much of the usage activity as portrayed in the logs can be explained and understood. The data triangulates. In other words, logs provide a sound indicator of use and information seeking behaviour, albeit in respect to just a part of the information seeking journey. Above all, the two sets of data confirm and explain the 'lite' form of information seeking behaviour observed in the logs, a sure sign of users trying to make their minds up in the face of massive choice as to what is relevant and can be trusted. After all, in a disintermediated, environment researchers are their own librarians, and while probably this was ever the case for senior researchers it is even truer now.

The data also confirms the marked preference for Google and Google Scholar as discovery tools among academics and researchers. They are as influenced in what they use or read by ease of access as the general public; and young researchers were most influenced. This would appear to be a significant change in behaviour. It does appear that researchers tend to use publisher platforms as warehouses, places to obtain the full text, but not to do much searching; searching will be undertaken in Google of a gateway site, such as PubMed Central. This, together with the fact that they tend to get much of their reading material from colleagues might well provide an explanation for the short visits so characteristically found in the usage logs of publisher and library websites. After all, if you know what you are looking for already, you are not going to dwell long in the publisher/library space. Google and Google Scholar are clearly the trusted discovery platforms and in this respect has usurped the library and publisher platform.

Also in respect to trustworthiness and quality:

- The sheer amount of activity associated with peer reviewed journals confirms loudly and clearly that they are highly sort and trusted information sources for science and the social sciences.
- However, while some journals are used very heavily, a large percentage are hardly used at all. Clearly this has to be a sign that many journal articles have very little interest and value at all. In fact, a recent French study showed the sheer scale of this, with the top 50% of journal titles generating 95% of total downloads, so the other 50% were pretty redundant (Boukacem-Zeghmouri et al., 2014). This is a sure sign of researchers voting with their mouse clicks or buttons.
- Usage data and factors are not seen to be a proxy for quality among academics and researchers. The fact that something is used heavily does not mean it is trustworthy, worthy of citing, for instance. Not all full-text views are equal, some might represent nothing more than a cursory inspection and rejection of an article, some might constitute a rapid recognition of relevance and download to read offline later, others might involve the reading of just a page or two of the article, and yet others might involve the reading of the whole article online (Nicholas et al., 2008).
- Abstracts are very heavily used in order to assess trustworthiness, so important are they that researchers are calling for them to be more informative and peer reviewed.

The big problem that faces us all, information professionals and publishers alike, is making sense of the avalanche of usage data that has become available as a consequence of the digital transition. This paper has gone some way to showing how this can be done. However, the digital usage evaluation methods that are the subject of this paper, which have served information science researchers well enough for so long are just not effective, economical or practical anymore. Trying to analyse usage by working from a dump of pre-existing data —the basis of the raw log method — does not yield ready results in the way it once did. The task of filtering out all the noise — bots and crawlers, fragments and frameworks, images and styles— and reconstructing a browsing history has become overwhelmingly complex and consequently unreliable. Fortunately, the user knowledge gap which is worryingly opening up as a consequence may be filled by the ubiquitous Google Analytics, although the jury is still out on this (Nicholas et al., 2014).

Acknowledgments

Some of the research reported in this paper was funded by the Alfred P. Sloan Foundation.

The authors would like to thank the contributions of other members of the Sloan funded research project, namely Carol Tenopir, Rachel Volentine, Suzie Allard, Kenneth Levine, Lisa Christian, Reid Boehm, Frances Nichols, and Rob Christensen from the University of Tennessee.

References

Boukacem-Zeghmouri, C. et al. (2014). Retour sur Investissement (ROI) de la consultation des revues électroniques en bibliothèques universitaires françaises: Approches bibliométrique et économétrique. Research report. Available at <http://hal.archives-ouvertes.fr> Retrieved 2014.02.20.

Carr, N. (2011). *The Shallows: What the Internet Is Doing to Our Brains*. New York, W.W Norton.

CIBER. (2011). E-journals: their use, value and impact - final report. Research Information Network. Available at <http://www.rin.ac.uk/our-work/communicating-and-disseminating-research/e-journals-their-use-value-and-impact> Retrieved 2014.02.20.

CIBER and University of Tennessee. (2014a). Trust project. Available at <http://ciber-research.eu/CIBER_projects.html> Retrieved 2014.02.20.

CIBER and University of Tennessee. (2014b). Trust and Authority in Scholarly Communications in the Light of the Digital Transition. Final Report. Available at <http://ciber-research.eu/download/20140115-Trust_Final_Report.pdf> Retrieved 2014.02.20.

COUNTER. (2014). Usage-based measures of journal impact and quality. Available at <http://www.projectcounter.org/usage_factor.html> Retrieved 2014.01.20.

Howard, J. (2012). JSTOR tests free, read-only access to some articles [blog post]. *The Chronicle of Higher Education: Wired Campus*. Available at <http://chronicle.com/blogs/wiredcampus/jstor-tests-free-read-only-access-to-some-articles/34

908> Retrieved 2013.02.25.

Jamali, H. R., & Asadi, S. (2010). Google and the scholar: The role of Google in scientists' information-seeking behaviour. *Online Information Review*, 34(2), 282-294. doi: 10.1108/14684521011036990

Nicholas, D. (2010). The behaviour of the researcher of the future (the 'Google generation'). *Art Libraries Journal*, 35(1), 18-21.

Nicholas, D. (2010). The virtual scholar: the hard and evidential truth. In *Digital Library Futures*. IFLA Publication Series. K.G. Saur Verlag, Munich, 23-32.

Nicholas, D., & Clark, D. (2012). Evidence of user behaviour: deep log analysis in Milena Dobreva, Andy O'Dwyer and Pierluigi Feliciati, Editors. *User studies for digital library development*. London: Facet, 85-94.

Nicholas, D., Clark, DJ., & Jamali, HR. (2014). Evaluating information seeking and use in the changing virtual world: the emerging role of Google Analytics. *Learned Publishing*, 27(3), *in progress.*

Nicholas, D., Clark, D., Rowlands, I., & Jamali, H.R. (2009). Online use and information seeking behaviour: Institutional and subject comparisons of UK researchers. *Journal of Information Science*, 35(6), 660-676. doi: 10.1177/0165551509338341

Nicholas, D., Huntington, P., & Jamali, H. R. (2007). The Use, Users, and Role of Abstracts in the Digital Scholarly Environment. *Journal of Academic Librarianship*, 33(4), 446-453. doi: 10.1016/j.acalib.2007.03.004

Nicholas, D., Huntington, P., Jamali, H. R., & Dobrowolski, T. (2008). "The Information-Seeking Behaviour of the Digital Consumer: Case Study the Virtual Scholar." In: Nicholas, D. and Rowlands, I., Eds., *Digital Consumers: Reshaping the Information Professions*. London: Facet Publishing, 113-158.

Nicholas, D., Huntington, P., Jamali, H. R., Rowlands, I., Dobrowolski, T., & Tenopir, C. (2008). Viewing and reading behaviour in a virtual environment: The full-text download and what can be read into it. *Aslib Proceedings*, 60(3), 185-198. doi: 10.1108/00012530810879079

Nicholas, D., Huntington, P., Tenopir, C., Jamali, H., & Dobrowolski, T. (2008). Viewing and reading behaviour in a virtual environment: the full-text download. *Aslib Proceedings*, 60(3), 186-198.

Nicholas, D., Huntington, P., Williams, P., & Dobrowolski, T. (2004). Re-appraising information seeking behaviour in a digital environment: bouncers, checkers, returnees and the like. *Journal of Documentation*, 60(1), Jan/Feb, 24-39.

Nicholas, D., & Rowlands, I. Editors. (2008). *Digital Consumers; reshaping the information professions*. London: Facet, 2008.

Nicholas, D., Rowlands, I., Huntington, P., Jamali, H. R., & Salazar, P. H. (2010). Diversity in the e-journal use and information-seeking behaviour of UK researchers. *Journal of Documentation*, 66(3), 409-433. doi: 10.1108/00220411011038476

Nicholas, D., Williams, P., & Rowlands, I. (2010). Researchers' e-journal use and information seeking behaviour. *Journal of Information Science*, 36(5), August, 494-516.

PLOS. (2014). Article-Level Metrics measure the dissemination and reach of published research articles. Available at <http://www.plos.org/innovation/article-level-metrics/> Retrieved 2014.02.20.

Research Information Network. (2009). Available at: E-journals: their use, value and impact.

<http://ciber-research.eu/download/20090401-RINejournals.pdf> Retrieved 2014.01.20.

Tenopir, C., Allard, S., Bates, B., Levine, K., King, D.W., Birch, B., Mays, R., & Caldwell, C. (2011). Perceived Value of Scholarly Articles. *Learned Publishing*, 24, 123-132.

Town, S. (2004). E-measures: a comprehensive waste of time? *Vine*, 34(4), 190-195.