If You Can't Retrieve it, Does it Exist? Accessibility of LIS Journals on the Internet

Tove Faber Frandsen

University of Southern Denmark, Campusvej 55, DK-5230 Odense M, Denmark.
Email: tofr@litcul.sdu.dk

Jeppe Nicolaisen

Royal School of Library and Information Science, Birketinget 6, DK-2300 Copenhagen S., Denmark.
Email: jni@iva.dk

Abstract: *Public access to the World Wide Web became widespread in the late 1990s and early 2000s, and today documents are frequently published on the Internet. Open access (OA) to the scientific literature has been found to be increasing as more and more producers and publishers of scientific literature make their publications available free of charge on the Internet. The paper puts forward that it may be argued that only scholarly Internet documents that are retrievable through the search engine Google Scholar (GS) can be said to exist. The degree of coverage of GS is thus an important issue. The paper reports the results of a study of 159 journals in the field of Library and Information Science and their degree of coverage in GS. Journals publishing many issues a year are not found to be more retrievable than journals with fewer issues. Non-English and OA journals tend to have a lower degree of retrievability. The tendency is found to be even stronger for journals that are both OA and non-English. OA and non-English journals are very heterogeneous groups and the variation in their degree of retrievability is found to be much higher than in the case of traditional, toll-access journals, which resemble each other more in relation to retrievability.*

Keywords: *Google Scholar; retrievability; open access.*

Introduction

There is a famous riddle in philosophy that goes something like this:

> *If a tree falls in a forest and no one is around to hear it, does it make a sound?*

The riddle can be traced back at least to George Berkeley (1685-1753) and his work "A Treatise Concerning the Principles of Human Knowledge". Raising questions, as it does, about observation and reality, the riddle has occupied many bright thinkers ever since. The basis of the riddle is, of course, whether something can exist without being perceived. We shall not attempt to solve the riddle here, but only mention it in passing as it resembles a similar riddle we have discovered in our own field of Library and Information Science (LIS):

> *If a document is published but nobody can find it, does it exist?*

Imagine, for instance, a misplaced book in a library collection. The classification code in the library catalog tells a user where to find the book, but the book is not there. It was mistakenly placed at the wrong bookshelf by the last user. Does the book exist then? Just like the riddle about the tree falling in the forest, this riddle concerns the question whether something can exist without being perceived. It is a metaphysical question, and the answer depends at least in part on whether one believes that an object is distinct from its properties or whether an object is merely its sense data.

Taking the question about existence to a more concrete level, a number of studies have shown that scholars and scientists tend to adopt easily accessible information resources to a larger extent than less easily accessible information resources. One example is a study of neuroscientists performed by Vibert et al. (2007). This study suggests that when scientists can access efficient and exhaustive BDI resources online, those resources quickly become their preferred way of getting work-related

information. This corresponds to the findings of Tenopir et al. (2005) who found that astronomers tend to select access means that are convenient whether in print, electronic, or both. Thus, the prediction put forward by Odlyzko (2002) that scholars, publishers and librarians will have to make even greater efforts to make their material easily accessible is as true today as it was almost ten years ago. In recent years there has been a major change in the ways scientific communication (both formal and informal) is disseminated by electronic means. At the same time, the information seeking behavior of scientists have changed, especially when it comes to the use of search engines and digitized resources apart from traditional journals (Meyer & Schroeder, 2009). The electronic accessibility of journals implies, among other things, that scientists now make fewer library visits. Web browsing and table of contents e-mail alerts are gradually replacing physical browsing, and searching has been found to be a popular option for keeping up to date with new developments (Olle & Borrego, 2010). However, like the misplaced book in the library collection, one may ask whether this open access (OA) literature really exists. Is 'putting something on the Internet' equal to an existence claim? We do not believe so. The Internet is humongous. Consequently, putting something on the Internet is in itself no existence claim. It's like the misplaced book in the library collection. There is no guarantee that the publication will ever be found again. Without going deeper into the metaphysical question about existence, we believe that a first requirement for the existence of an Internet document is that it is retrievable. This, of course, raises a question about what it means to be retrievable on the Internet. All documents on the Internet have a unique URL. In principle, all documents on the Internet are thus retrievable – one only needs to know the correct URL. Again, this compares to the misplaced book in the library collection. One could find it if one knew where to look. Thus, a unique URL is not enough. Limiting the discussion to scientific documents, we will argue that a document is retrievable if it can be found by searching a proper search engine. What, then, is the proper search engine for scientific documents on the Internet? We believe it is Google Scholar (GS) (scholar.google.com). This belief is based in part on its presumed preference among scientists (Jamali & Asadi, 2010) and in part on reports stating that GS ensures a higher precision in the search results compared to the comparable search engines due to its more narrow definition of scholarly literature:

> Similar in approach, but broader and less specific in scope than Google Scholar, the scientific search engine Scirus (www.scirus.com) searches, according to information they provide, approximately 300 million science-specific web pages. In addition to scientific documents from Elsevier (ScienceDirect server, see www.sciencedirect.com/) freely accessible documents are provided, many from public web servers at academic institutions. Among these are, for example, documents placed by students that do not fulfill scientific criteria such as peer review, which often lead to their exclusion in searches. In our experience there is more than a negligible fraction of records from Google Scholar non-academic web spaces in the Scirus index. Scirus' coverage of purely scientific sources in addition to Elsevier's ScienceDirect full-text collection is low by comparison (compare the selection of hosts in the Scirus advanced search interface, http://scirus.com/srsapp/advanced/). What Scirus declares as the "rest of the scientific web" is too general, non-specifically filtered and makes up the majority of hits in any query. (Mayr & Walter, 2007, pp. 815-816)

Consequently, if one wants to find scientific documents on the Internet, GS is currently the place to go and search.[1] But how well does GS perform? To what extent does it cover the scientific literature on the Internet?

A study by Chen (2010) finds that GS covers as much as 98 to 100 percent of scholarly journals from both publicly accessible Web contents and from subscription-based databases that GS partners with. Lewandowski (2010) measures the coverage of GS for Library and Information Science (LIS) journal literature as identified by a list of core LIS journals from a study by Schlögl and Petschnig (2005). He checked every article from 35 major LIS journals from the years 2004 to 2006 for availability in GS. The information on type of availability (whether a certain article was available as a PDF for a fee, as a free PDF or as a preprint) was analyzed and divided by the type of publisher. However, the study does not include open access as a specific characteristic of the journals in the analysis. Lewandowski (2010, p. 257) states:

---

[1] We acknowledge that there are also limitations and problems with GS (see e.g. Jacso (2009; 2010) and http://en.wikipedia.org/wiki/Google_Scholar#Limitations_and_criticism), but to our knowledge no other search engine currently matches up. In some fields Google Scholar performs better in terms of both recall and precision than most of the subscription databases when submitting simple keyword queries (Walters, 2009) whereas in others Google Scholar matches the results of a subscription database in terms of recall but not precision (Anders & Evans, 2010; Freeman et al., 2009).

> The numbers for the preprints are somewhat disappointing as there are high hopes for open access (OA) and the willingness of authors to make their work available through OA. Particularly for the LIS profession with its many OA promoters, the numbers seem to be very low.

The conclusion by Lewandowski is supported by Way (2010) who concludes that "the archiving of articles is not a regular practice in the field" (p. 302). Way (2010) measures the open access availability of LIS research in GS by searching for articles from 20 top LIS journals. The study does not focus on open access journals but concludes that GS is an effective search tool for retrieving LIS articles.

Open access can also be in the form of open access journals and, consequently, we would like to address the issue about GS coverage in the present study by including a substantial number of open access journals. The paper is, of course, a coverage study of an Internet search engine, but as we have tried to argue, our topic also concerns deeper philosophical questions.

Methods

With this study we want to investigate whether the open access LIS journal literature is retrievable in GS regardless of the form of publication indexed (e.g. preprint or publisher version). First of all we need to draw a sample of LIS literature on the basis of some formal characteristics. To generate our dataset we used the list of LIS journals available on Journal Citation Reports (JCR) and Directory of Open Access Journals (DOAJ). The former database list includes 66 journals and the latter 117. Excluding duplicates, our dataset for the current study consisted of 177 journals (see Table 1 for the journal list). For the present study all document types are included. The number of publications published in 2009 was checked on the journals' web sites. No distinction was made between various publication types. Obviously some publication types report research to a greater extent than others. Only publications such as call for papers or author instructions were excluded from the study. In the discussion and conclusion section we will return to how future work could strengthen the data set.

Now we need to determine the variables to be included in the study.

The dependent variable is number of retrievable documents in GS. An operationalization of the concepts of "visibility", "retrievable" or "indexed by GS" is necessary. A publication can hardly be said to be visible and retrievable, if it only appears in GS with bibliographic information in a bibliography. The number of subject access points is relatively impeded and we will not gain access to the full publication, an abstract or subject headings. In this study a document is determined to be retrievable from GS if the document is indexed with subject access points beyond the bibliographic reference, i.e. abstract, full text or subject headings. In some cases toll access prevents us from gaining access to the full document although it is indexed in GS. Each journal was checked in GS for the publication year of 2009. Noruzi (2005) summarizes the search techniques available through GS. Peter Jacso (2005; 2008) argues that the size of the database has increased substantially from 2005 to 2008 and GS primarily lacks satisfactory levels of correct assignment of journal names and authors which causes difficulties in the retrieval of articles. However, random examination indicates that the problem is atypical in this case. In this case the queries submitted were based on journal name and publication year using the advanced scholar search formular.

Furthermore, we need to determine the appropriate independent variables. Forms of open access could be an independent variable. Lewandowski (2010) finds that the availability of preprint versions of articles is somewhat disappointing as only a low number of articles is available in this form. The author finds that it is not caused by the indexing by GS but it is rather a result of the lack of self-archived publications by LIS authors. This leads the author to conclude that "while many researchers and practitioners in the LIS field are advocates of open access, when it comes to advancing open access through depositing preprints, their support is limited" (Lewandowski, 2010, p. 260). Consequently, we choose not to separate availability in forms. We focus on whether the articles are indexed (regardless of whether there is open access to a full text version).

Secondly, language needs to be considered as an independent variable. Lewandowski (2007, 2010) finds that a much higher percentage of English-language LIS articles is available in GS. However, the non-English journals are typically published mainly by small publishers and, for some, no online version is available, therefore there is a lower coverage rate. The author concludes: "The results do

not indicate a language or country bias towards English-language articles" (Lewandowski, 2010, p. 260). We would like to extend the study of non-English journals to other languages than German and to non-English journals with an online version. Consequently, language is added as a variable in the data set.

The methods of our study followed this procedure: Each journal was entered into the data set with ISSN to uniquely identify the journal, journal name, a dummy variable for language (non-English=1 if journal included publications in other languages than English, otherwise 0), dummy variable for open access (OA=1 for open access, otherwise 0), number of publications in 2009 (according to the journal web sites or alternatively the citation indexes) and number of publications indexed in GS. On the basis of that data we can calculate the share of publications indexed in GS.

Results

First of all, we need to consider a potential correlation between number of publications published and the share of these available through GS. One might argue that small journals (i.e., publishing fewer issues annually) find it difficult to attract attention generally and more specifically by GS whereas large journals (i.e., publishing more issues annually) are generally more visible. Figure 1 illustrates the relation between number of publications and the share available in GS. One outlier has been excluded from the figure as it distorted the picture significantly.

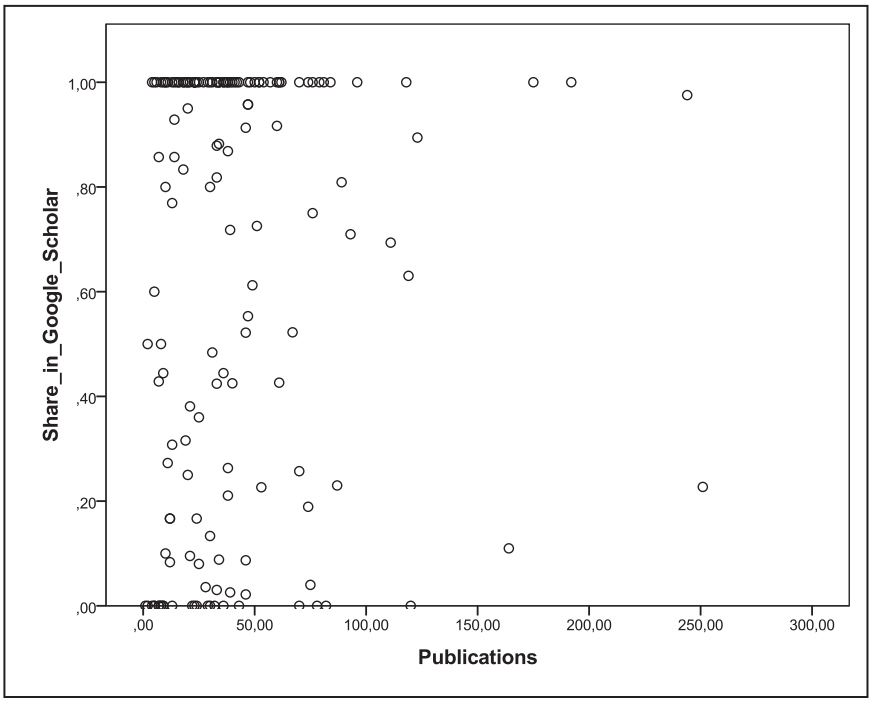

Figure 1. Share in GS and number of issues published annually by the journal

From the figure it is clear that such correlation cannot be found. Statistical analyses confirm the picture depicted in the figure. Linear regressions (with and without the outlier) with share in GS as the dependent variable and publications as the independent variable reveal p-values of the independent variable ranging from .25 and .29. Consequently, we move on with the analyses.

Table 1 provides an overview of the mean and median values. Mean is typically used to describe the central tendency of a set of data that does not have extreme values (outliers). On the other hand, the median is typically used to describe the central tendency of a set of data that does have extreme values as it is not affected as strongly as the mean by outliers. In this case the values range from 0 to 1 and there are no outliers (see Figure 1 and Table 1). Consequently, the mean is the appropriate measure to describe the middle.

Table 1. Overview of results

Share in Google Scholar							
		N	**Mean**	**Median**	**Percentile 05**	**Percentile 25**	**Percentile 75**
Open_access	0	62	.79	1.00	.10	.63	1.00
	1	97	.52	.50	.00	.04	1.00
Non_English	0	108	.68	.88	.00	.40	1.00
	1	51	.51	.26	.00	.00	1.00
OA_Non_English	0	114	.68	.88	.00	.36	1.00
	1	45	.50	.23	.00	.00	1.00

In the following figures we use the mean to analyse differences in coverage. Figure 2 illustrates the visibility in GS of OA journals and traditional toll access journals. The error bars represent the 95% cent confidence interval. We can clearly see that although there is some variance among journals within the two categories of journals, traditional toll access journals have higher coverage rates in GS than OA journals. The former have a mean coverage of 79% cent whereas the latter is 52%.

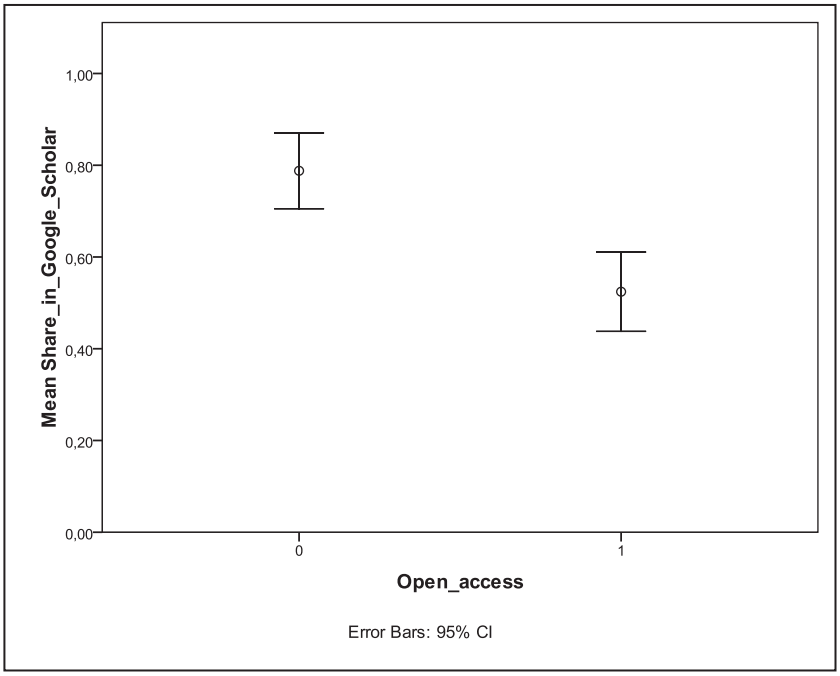

Figure 2. Mean share indexed in GS and open access status. The error bars represent the 95% cent confidence interval

Figure 3 illustrates the visibility in GS of journals that are published exclusively in English and journals that are published in English as well as in other languages. Journals published exclusively in English have higher visibility in GS as their share in GS is 17 percentage points higher than those published in other languages than English. The latter group seems to be a more heterogeneous group as their variance is greater.

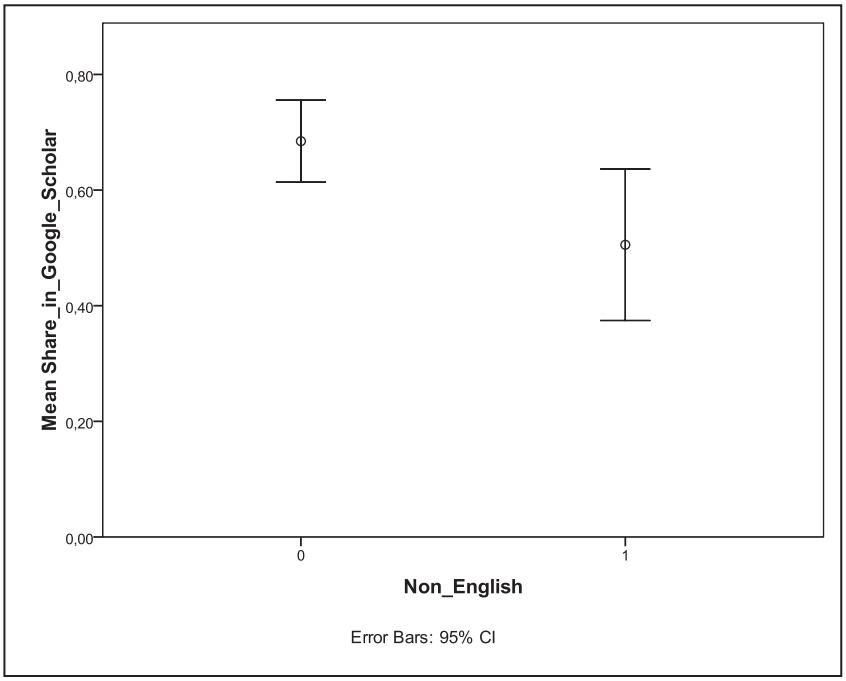

Figure 3. Mean share indexed in GS and language. The error bars represent the 95% cent confidence interval

Figure 4 illustrates the visibility in GS of OA journals that are published in English as well as in other languages compared to toll access journals that are published exclusively in English.

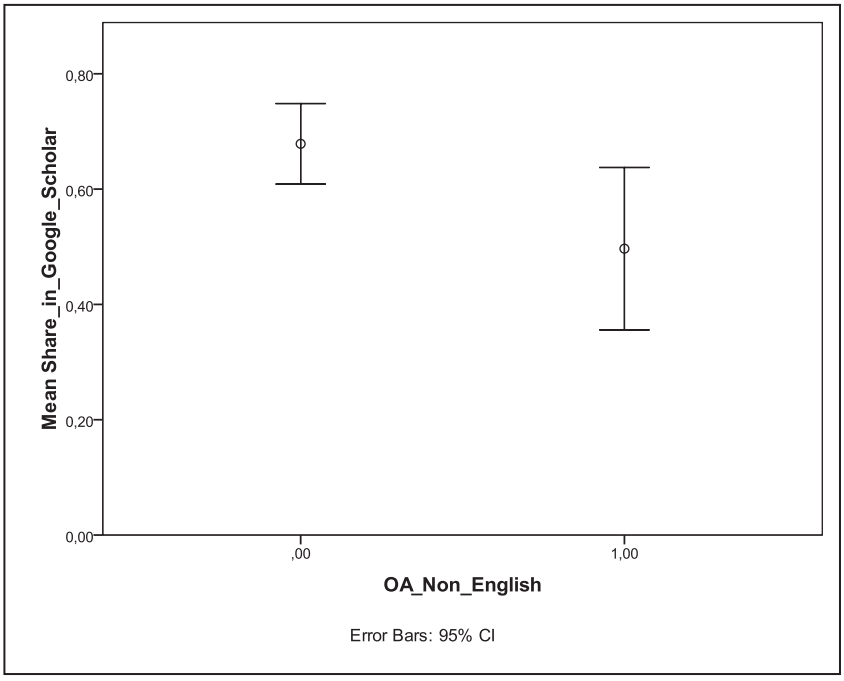

Figure 4. Mean share indexed in GS and OA as well as non-English. The error bars represent the 95% cent confidence interval.

The open access journals including publications not in English are characterised by an even lower share of visibility in GS. The remaining journals of the data set cannot be characterised differently in terms of mean than the non-English journals. Their share is, however, 11 percentage points lower than that of the toll access journals.

Discussion and Conclusion

Before engaging with the discussion of the implications of our results, we will briefly summarize our findings. First of all, large journals (i.e., journals with many issues published every year) are not more retrievable in GS than smaller journals. Non-English and OA journals tend to have a lower degree of retrievability in GS. The tendency is even stronger for journals that are both OA and non-English. In addition, OA and non-English journals are very heterogeneous groups and there is much more variation in their degree of retrievability than in the case of traditional, toll-access journals, which resemble each other more in relation to retrievability.

As we have tried to make it clear throughout the paper, OA is also a question about retrievability, but there need not be a correlation between open access and retrievability. Putting something on the Internet free of charge is generous, but if nobody is able to locate and access it, it will hardly make a difference.

The present work could be extended and strengthened by including more journals in the study. The present study casts light on the relationship between open access and retrievability for open access and journals with enough international impact to be included in JCR. A number of journals with different characteristics than the journals in this study such as toll access journals from developing countries could help depict a richer picture of the relationship between open access and retrievability.

Using GS as our index of retrievability, we have found large differences in the retrievability among OA journals. The responsibility for this lies with the editors. It seems that some OA journals could benefit from a more active approach to having their content indexed by GS. There are technical inclusion guidelines as well as an inclusion request form available at the GS website (http://scholar.google.com/intl/en/scholar/publishers.html#faq1). This study stresses that editors of OA journals should be aware of the importance to comply with these guidelines because in a way: *If you can't retrieve it, it doesn't exist!*

References

Anders, M.E. & Evans, D.P. (2010). Comparison of PubMed and Google Scholar Literature Searches. *Respiratory Care*, 55(5): 578-583.

Chen, X.T. (2010). Google Scholar's dramatic coverage improvement five years after debut. *Serials Review*, 36(4): 221-226.

Freeman, M.K., Lauderdale, S.A., Kendrach, M.G., & Woolley, T.W. (2009). Google Scholar versus PubMed in locating primary literature to answer drug-related questions. *Annals of Pharmacotherapy*, 43(3): 478-484.

Jacso, P. (2005). Google Scholar: The pros and cons. *Online Information Review*, 29(2): 208-14.

Jacso, P. (2008). Google Scholar revisited. *Online Information Review*, 32(1): 102-14.

Jacso, P. (2009). Google Scholar's ghost authors. *Library Journal*, 134(18): 26-27.

Jacso, P. (2010). Metadata mega mess in Google Scholar. *Online Information Review*, 34(1): 175-191.

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

Lewandowski, D. (2007). Nachweis deutschsprachiger bibliotheks- und informationswissenschaftlicher Aufsätze in Google Scholar. *Information Wissenschaft und Praxis*, 58(3): 165-168.

Lewandowski, D. (2010). Google Scholar as a tool for discovering journal articles in library and information science. *Online Information Review*, 34(2): 250-262.

Mayr, P. & Walter, A.-K. (2007). An exploratory study of Google Scholar. *Online Information Review*, 31(6): 814-830.

Meyer, E.T. & Schroeder, R. (2009). The world wide web of research and access to knowledge. *Knowledge Management Research & Practice*, 7(3): 218-233

Noruzi, A. (2005). Google Scholar: The new generation of citation indexes. *Libri*, 55(4): 170-180.

Odlyzko, A. (2002). The rapid evolution of scholarly communication. *Learned Publishing*, 15(1): 7-19.

Olle, C. & Borrego, A. (2010). A qualitative study of the impact of electronic journals on scholarly information behavior. *Library & Information Science Research*, 32(3): 221-228.

Schloegl, C. & Petschnig, W. (2005). Library and information science journals: an editor survey. *Library Collections, Acquisitions, & Technical Services*, 29(1): 4-32.

Tenopir, C., King, D.W., Boyce, P., Grayson, M., & Paulson, K-L. (2005). Relying on electronic journals: Reading patterns of astronomers. *Journal of the American Society for Information Science and Technology*, 56(8): 786-802.

Vibert, N., Rouet, J-F., Ros, C., Ramond, M., & Deshoullieres, B. (2007). The use of online electronic information resources in scientific research: The case of neuroscience. *Library & Information Science Research*, 29(4): 508-532.

Walters, W. H. (2009). Google Scholar search performance: Comparative recall and precision. *Portal-Libraries and the Academy*, 9(1): 5-24.

Way, D. (2010). The open access availability of library and information science literature. *College & Research Libraries*, 71(4): 302-309.