Available online at www.sciencedirect.com
ScienceDirect
ELSEVIER
Information Processing and Management 43 (2007) 1121–1132
INFORMATION PROCESSING & MANAGEMENT
www.elsevier.com/locate/infoproman

Google Scholar coverage of a multidisciplinary field

William H. Walters *

Helen A. Ganser Library, Millersville University, Millersville, PA 17551-0302, USA

Received 10 July 2006; received in revised form 18 August 2006; accepted 19 August 2006
Available online 23 October 2006

Abstract

This paper evaluates the content of Google Scholar and seven other databases (Academic Search Elite, AgeLine, ArticleFirst, GEOBASE, POPLINE, Social Sciences Abstracts, and Social Sciences Citation Index) within the multidisciplinary subject area of later-life migration. Each database is evaluated with reference to a set of 155 core articles selected in advance—the most important studies of later-life migration published from 1990 to 2000. Of the eight databases, Google Scholar indexes the greatest number of core articles (93%) and provides the most uniform publisher and date coverage. It covers 27% more core articles than the second-ranked database (SSCI) and 2.4 times as many as the lowest-ranked database (GEOBASE). At the same time, a substantial proportion of the citations provided by Google Scholar are incomplete (32%) or presented without abstracts (33%).
© 2006 Elsevier Ltd. All rights reserved.

Keywords: Assessment; Bibliographic database; Comparison; Index; Retrieval

1. Introduction

Google Scholar (Google, 2006), released in November 2004, immediately attracted the attention of librarians and information professionals due to its potential as a free online search engine for scholarly literature in a wide range of subject areas (Gorman, 2006; Mullen & Hartman, 2006; Myhill, 2005; Peek, 2005; Tenopir, 2005). Unlike the regular version of Google, which searches all publicly available Web pages, Google Scholar (GS) gets its records from three sources (Google, 2005; O'Leary, 2005). First, GS uses a proprietary algorithm to identify those publicly accessible Web documents that "look scholarly" in their content and format—journal articles, preprints, dissertations, and technical reports, for example. The materials identified in this way include both full-text documents and citations with abstracts. Second, Google Scholar adds content provided by its partners—journal publishers, scholarly societies, database vendors, and academic institutions. Much of this material is taken, with permission, from university intranets or from publishers' restricted-access sites. This content is not accessible through the regular Google interface and consists chiefly of citations with

* Tel.: +1 717 871 2063; fax: +1 717 872 3854.
E-mail address: william.walters@millersville.edu

0306-4573/$ - see front matter © 2006 Elsevier Ltd. All rights reserved.
doi:10.1016/j.ipm.2006.08.006

abstracts. (All GS partners agree to make abstracts of their documents freely available to the public through Google Scholar. In most cases, however, the full text is available only to subscribers.) Third, Google Scholar extracts citations from the reference lists of documents found through the first two methods, listing them as separate search results. Items identified through this third procedure are available only as citations without abstracts.

For each Google Scholar search, the records with the highest relevance scores are presented first in the list of search results. GS computes relevance scores on the basis of several factors including citedness—the number of papers within the GS database that have cited each document. The citation counts reported by Google Scholar are not entirely reliable, however (Bar-Ilan, 2006; Jacsó, 2005c, 2006).

2. Aims and objectives

This study presents a comparative assessment of Google Scholar and seven other databases, focusing specifically on each database's coverage of 155 core articles representing the most important papers on later-life migration published from 1990 to 2000. Later-life migration is a multidisciplinary field of interest that encompasses elderly migration, retirement migration, and related forms of geographic mobility, both seasonal and permanent. The subject is specific enough to be covered in undergraduate term papers yet broad enough to be recognized as a distinct area of inquiry.

Although several authors have investigated Google Scholar's effectiveness in retrieving useful information, this is the first study to assess its coverage of documents that have been identified beforehand based on their importance within a particular area of inquiry. This approach—searching for documents that appear on a list of items known to be relevant—is consistent with the procedure adopted by Griffith, White, Drott, and Saye (1986) in their assessment of database coverage in the medical behavioral sciences. The standard of relevance employed here is not "Is this document useful?" but "Is this document an important contribution to the field? Is it a paper that ought to be read by those who want to develop a solid grasp of the subject?" Admittedly, this standard does not account for the needs or objectives of individual database users. It does, however, incorporate two kinds of information not usually considered in database assessment studies: the content of each paper (which provides a better guide to relevance than the abstract alone) and an understanding of the broader context (the universe of scholarly information within the field of interest). While most database comparison studies have relied on the assumption that relevant papers can be identified through subject searching alone (Tenopir, 1982), many bibliographers would argue that the most authoritative and reliable assessments require a careful reading of every potentially relevant paper.

This analysis deals with database content rather than search effectiveness. Specifically, it examines the relative number of core articles indexed by each database, and the extent to which the core articles covered by the other seven databases can be found in Google Scholar as well. The study also evaluates Google Scholar's publisher and date coverage, and examines the sources of its citations, abstracts, and full-text links. Although the analysis is largely exploratory, it does focus on a few basic questions: Is Google Scholar biased in favor of those publishers with which it has cooperative arrangements for the provision of abstracts and full-text links? Does it provide less comprehensive coverage of those articles that are not available online—those published before the most rapid growth of the Internet (ca. 1996)? With regard to overall coverage, there are at least two possibilities. On the one hand, GS might be expected to provide especially good coverage of the literature due to its size and interdisciplinary scope. On the other hand, it might provide especially poor coverage due to the absence of systematic mechanisms that would ensure comprehensive indexing of key journals or topics within any particular subject area.

3. Previous research

Reviews of Google Scholar have tended to focus on the search interface rather than the database content. Specifically, quite a few authors have noted the apparent deficiencies of GS in comparison with other bibliographic databases. There is no controlled vocabulary for subject terms, no authority files for author names or journal titles, no way to limit the results to those records with full-text links, no way to sort the results other than by Google's own assessment of relevance, and no way to save or export citations or abstracts. Moreover, the Google staff have been unwilling to provide much information about Google Scholar's selection algorithms, sorting (relevance) algorithms, content sources, or organizational partnerships. Perhaps because of these difficulties, many librarians have been hesitant to market GS to their patrons. Of the 113 university libraries surveyed by Mullen and Hartman (2006), only 24% include Google Scholar in their alphabetical lists of databases. Only 5% include it in their public-access catalogs.

3.1. Subject-based assessments

Several authors have evaluated Google Scholar's coverage of particular topics. Wleklinski, searching for information on the political scientist Ithiel de Sola Pool, found relatively many citations to non-scholarly and non-authoritative sources. She asked, "Could I have written a credible research paper from the information Google Scholar returned? Of course not. Did it lead me to other sources? Definitely. Most importantly, it gave me a good first-start overview on the subject" (Wleklinski, 2005, p. 24). Gardner and Eng (2005) reached a similar conclusion when searching for papers on homeschooling in Google Scholar, ERIC, PsycINFO, and Social Sciences Citation Index. They compared the first 100 results from each database, noting that Google Scholar retrieved far more records than the others, that it failed to find the most recent articles, and that there was little overlap (2–8%) between the first 100 Google Scholar records and the first 100 records found in any of the other databases. "There is more variety in Google Scholar and a higher number of results, but they are not necessarily as scholarly or relevant... Since there is less quality control, some results are of questionable research value" (Gardner & Eng, 2005, p. 43).

Jones (2005) assessed the performance of 10 biological and multidisciplinary databases by searching for papers on Nodilittorina, a type of periwinkle. Because BIOSIS is the most comprehensive database in biology, she looked specifically at the proportion of BIOSIS records that were retrieved by each of the other databases. Overall (counting results from 1969 to the present), Google Scholar returned citations for 27% of the 110 documents indexed in BIOSIS—far more than any other database in the comparison group. GS also provided records for 102 items not found in BIOSIS, including many from non-English journals. For recent documents (1996 to 2005), Google Scholar's performance was even better; GS retrieved 56% of the 39 BIOSIS citations plus 91 additional records.

In the most comprehensive assessment of Google Scholar's coverage, Neuhaus, Neuhaus, Asher, and Wrede (2006) generated a random sample of 50 journal articles from each of 47 databases, then calculated the proportion of those articles that were indexed in Google Scholar. Overall, citations for 60% of the sampled articles were found in GS. The results varied dramatically by subject area, however. Google Scholar provided 76% coverage (on average) in the natural sciences but only 41% coverage in education, 39% coverage in the social sciences, and 10% coverage in the humanities.

3.2. Publisher- and journal-based assessments

Many of the studies comparing GS with publishers' own databases have reported deficiencies in Google Scholar's coverage. The most common complaint is that GS fails to provide complete coverage of the content made available by its partner organizations. In December 2004, for example, Google Scholar provided citations for only half the recent articles published by Blackwell, one-third of the recent articles published by Wiley, and one-fifth of the recent articles published by Taylor and Francis (Hamaker & Spry, 2005). More recently, GS was found to include less than half the bibliographic records available at the *Nature* Web site and only 30% of the records available at the *Science* Web site (Giustini & Barsky, 2005; Jacsó, 2005d). Notess (2005), searching for papers on protonation alkylation in the summer of 2005, found approximately 1820 hits in Google Scholar but 21,685 hits in the proprietary American Chemical Society database.

Jacsó (2004, 2005e) compared GS with the digital archives of five major publishers, reporting that Google Scholar returned consistently fewer search results than the publishers' own databases—only 43% as many Institute of Physics hits as IoP's own Web site, for example. Moreover, the number of hits in GS is potentially far greater than the number of documents retrieved, since any one document may be represented by several bibliographic records taken from various sources. According to Jacsó (2004), duplicate and triplicate records are not uncommon.

Another criticism is that Google Scholar is not updated often enough. Burright (2006), Tennant (2005a, 2005b) and Vine (2006) discovered that new articles may not show up in GS for months after their appearance on publishers' Web sites. Likewise, Neuhaus et al. (2006) reported typical lag times of 12–15 weeks. In contrast, Hamaker and Spry (2005, p. 71) concluded that "Google Scholar did an excellent job of finding recent publications" within the set of Blackwell nursing journals.

3.3. Citation-based assessments

A significant number of Google Scholar citations are extracted from the reference lists of other documents through an automated procedure unique to GS. Unfortunately, these citations are not always complete or accurate (Tennant, 2005a). For example, Google Scholar recently indexed a physics paper by "Blegdamsvej" and "Copenhagen"—street and city names found in the address of the second author (Ojala, 2005). In their search for information on homeschooling, Gardner and Eng (2005) found that "H. Schooling" was incorrectly listed as the author of an article.

Several authors have evaluated Google Scholar's effectiveness as a citation index. In an analysis of papers published in the *Asian Pacific Journal of Allergy and Immunology* from 1983 to 2004, Jacsó (2005b) found that Science Citation Index identified nearly 2.3 times as many "cited by" entries as Google Scholar. However, a similar study by Bauer and Bakkalbasi (2005) revealed that GS provides more comprehensive results for more recent articles. Specifically, Bauer and Bakkalbasi compared the citation counts reported in three databases for papers published in the *Journal of the American Society for Information Science* in 1985 and 2000. For 1985 articles, the citation counts reported in Social Sciences Citation Index (SSCI) were higher than those reported in GS and Scopus. For 2000 articles, however, the citation counts reported in GS were higher than those reported in SSCI and Scopus.

Bar-Ilan (2006) evaluated the characteristics of three citation databases (Google Scholar, Citeseer, and Science Citation Index) in a comprehensive analysis of papers associated with the mathematician and computer scientist Michael O. Rabin. She found that GS performed reasonably well overall but failed to cite many of the papers that had not been made available online.

Although no citation database is entirely reliable, the absence of authority files for authors or journals seriously limits the citation-tracking capabilities of Google Scholar (Bar-Ilan, 2006; Jacsó, 2005a). Jacsó (2006) describes some of the citation-related inconsistencies found in the database, including one case in which a 2006 journal article was cited (according to GS) by several articles that had been published in the 1970s.

4. Methods

Database searching, journal browsing, and citation tracing were used to identify over 500 papers on later-life migration published from January 1990 to December 2000. Of those papers, 181 items (including 155 journal articles) were selected as the most important studies of the decade. Specifically, each paper was read and evaluated on the basis of six criteria: subject matter, importance of findings, innovativeness of methods or approach, availability of other literature on the topic, accessibility of content (potential readership), and accessibility of the paper itself. (See Walters, 2002, for details.) The evaluative process was conducted without knowledge of the databases in which each study was indexed, and the papers were reviewed in chronological order. The evaluations reflect the judgment of a single reviewer, however, and were not checked by others for reliability. The 155 journal articles identified through this procedure comprise the set of *core articles* used in this investigation. They include all the works cited by Walters (2002) except for those published before 1990 (40 items); those published after 2000 (1 item); those published as books, book chapters, or dissertations (26 additional items); and those that are primarily bibliographic or editorial (Findlay, 1991, 1992, 1993; Nash, 1994a, 1994b, 1996; Ogden, 1998, 2000; O'Lear, 1997; Warnes, 1990).

For each core article, title keyword searches were carried out in Google Scholar, Academic Search Elite, AgeLine, ArticleFirst, GEOBASE, POPLINE, Social Sciences Abstracts, and Social Sciences Citation Index. Except for Google Scholar, these databases are among those that have been found to provide good coverage of later-life migration (Walters & Wilder, 2003). AgeLine and POPLINE, both freely available to the public, were accessed through the Web sites of the AARP and the Johns Hopkins School of Public Health. All title searches were conducted in April and May of 2006. The AND operator was not used between words. If a colon or other punctuation appeared within the title, the search was conducted both with and without the punctuation.

The first 30 records returned by each search were examined in an effort to determine whether the core article was represented in the database as a citation, an abstract, and/or a link to a full-text document. Because some articles were represented by multiple records in Google Scholar, there were a few cases in which the citation appeared early in the list of search results but the abstract or full-text link appeared only later. Ninety-three percent of the core article citations found in Google Scholar appeared in the #1 position, as did 94% of the abstracts and 88% of the full-text links. Only 3% of the citations, 2% of the abstracts, and 6% of the full-text links were found beyond the fourth record in the list of search results.

The source of each citation, abstract, or full-text link was also recorded. (Only brief citations are presented on the Google Scholar results pages. Most brief citations include a link to a non-Google site at which complete bibliographic information and an abstract can be found.)

5. Results

5.1. General findings

Among the eight databases, Google Scholar provides the most comprehensive citation coverage of core articles in the field of later-life migration. (See Table 1.) Specifically, GS includes bibliographic records for 144 of the 155 core articles (93%). It covers 27% more core articles than SSCI and at least twice as many as each of the three disciplinary databases (AgeLine, POPLINE, and GEOBASE). These data show that Google Scholar indexes not just a large number of documents, but a large number of high-quality research papers.

Google Scholar's provision of abstracts is not quite as impressive, however. While five of the eight databases provide abstracts with virtually all their bibliographic records, GS provides abstracts for only 97 of the 144 core articles retrieved (67%). Further analysis reveals that there is a clear pattern to Google Scholar's abstract coverage. On the positive side, abstracts are included with all but two of the 99 citations representing publicly accessible Web pages or partner-provided content. At the same time, however, abstracts are not available for any of the 45 citations extracted from the reference lists of other documents. This is understandable, given the source of those citations, although it does put Google Scholar at a disadvantage relative to most of the other databases.

Google Scholar, through its partners, provides full-text links to 36 of the 155 core articles (23%). Full text is freely available for only one core article, however. Otherwise, full-text documents are available only to those individuals and institutions that have paid for subscriptions or article-level access.

Several of the idiosyncrasies noted in earlier studies were apparent during this investigation as well. For one thing, the GS results pages list no more than four authors for any one article and no more than 99 characters

Table 1
Citation and abstract coverage of the core literature in Google Scholar and seven other databases

	GS	SSCI	Article-First	Soc. Sci. Abs.	Age-Line	POP-LINE	Acad. Srch. Elite	GEO-BASE
Number of core articles	155	155	155	155	155	155	155	155
Number of core articles for which records (citations) are provided	144	113	94	86	73	67	62	60
Number of core articles for which abstracts are provided	97	113	0	36	73	67	60	60
Percentage of core articles for which records (citations) are provided	93	73	61	55	47	43	40	39
Percentage of records (citations) for which abstracts are provided	67	100	0	42	100	100	97	100

of any article title. Consequently, four of the brief citations to core articles were missing authors' names, and 24 had incomplete titles. (Fortunately, the entire title is searchable—even the part that's not displayed.) This analysis revealed one citation error (a misspelled name) and one link that led to the wrong destination page at the partner's Web site. Potentially more important, however, is the fact that not every core article has a bibliographic record in Google Scholar. After all, we might expect 100% coverage since every one of the 155 core articles has been cited in a literature review that was indexed by GS. If Google Scholar's citation extraction algorithms had been working systematically and reliably, each of the 155 citations would have been extracted from the reference list of Walters (2002) and made into a citation-only record.

Two of the characteristics of GS noted in earlier research were not evident here. While Jacsó (2004) found a significant number of duplicate and triplicate records, only a few core articles in the field of later-life migration were represented more than once in Google Scholar (within the first 30 search results). Likewise, this analysis revealed only two cases in which a single study was represented in more than one form. The publication of results in multiple venues—first in a conference volume, then in a journal, for instance—is

Table 2
Publisher coverage of Google Scholar and seven other databases

	GS	SSCI	ArticleFirst	Soc. Sci. Abs.	AgeLine	POPLINE	Acad. Srch. Elite	GEO-BASE
Number of core articles for which records (citations) are provided								
Sage (35)[a]	29	24	25	19	32	8	15	0
Blackwell (34)	34	32	19	22	7	23	19	25
Gerontological Society of America (24)	23	24	16	24	22	6	11	0
Rural Sociological Society (8)	8	8	8	8	0	3	6	7
Springer (7)	7	6	6	0	4	3	1	4
Hodder Arnold (5)	5	5	5	0	1	2	2	3
Pion (5)	4	5	3	5	1	3	3	5
Wiley (5)	5	0	0	0	1	3	1	5
Community Development Society (3)	3	0	1	0	0	1	0	0
Haworth (3)	2	0	1	1	3	2	1	2
Population Association of America (3)	3	3	3	3	1	3	1	2
US Department of Agriculture (3)	2	0	0	0	0	0	0	0
American Economic Association (2)	2	2	0	2	0	1	0	0
Arizona Business (2)	1	0	0	0	0	0	0	0
16 Publishers, each with 1 article (16)[b]	16	4	7	2	1	9	2	7
Percentage of core articles for which records (citations) are provided								
Sage (35)[a]	83	69	71	54	91	23	43	0
Blackwell (34)	100	94	56	65	21	68	56	74
Gerontological Society of America (24)	96	100	67	100	92	25	46	0
Rural Sociological Society (8)	100	100	100	100	0	38	75	88
Springer (7)	100	86	86	0	57	43	14	57
Hodder Arnold (5)	100	100	100	0	20	40	40	60
Pion (5)	80	100	60	100	20	60	60	100
Wiley (5)	100	0	0	0	20	60	20	100
Community Development Society (3)	100	0	33	0	0	33	0	0
Haworth (3)	67	0	33	33	100	67	33	67
Population Association of America (3)	100	100	100	100	33	100	33	67
US Department of Agriculture (3)	67	0	0	0	0	0	0	0
American Economic Association (2)	100	100	0	100	0	50	0	0
Arizona Business (2)	50	0	0	0	0	0	0	0
16 Publishers, each with 1 article (16)[b]	100	25	44	13	6	56	13	44
Standard deviation	16	47	38	45	37	26	24	40

[a] Numbers in parentheses indicate the number of core articles from each publisher.

[b] These include the American Geographical Society, the Association for Humanist Sociology, azcentral.com, Bellwether Publishing, Cambridge University Press, Crain Communications, Elsevier, Gamma Theta Upsilon, the Geographical Association, Greenwood Press, the International Association for Social Science Information Service and Technology, IOS Press, Lund University, Oklahoma State University, Taylor & Francis, and Western Michigan University.

apparently common among computer scientists (Bar-Ilan, 2006) but less common among those who study later-life migration.

5.2. Publisher coverage

Because Google Scholar relies on partnerships with individual publishers, we might expect relatively high variability in its publisher coverage. Some publishers might be represented fully while others might not be represented at all. Table 2 shows that this idea is not supported, however. Within the set of core articles on later-life migration, GS provides more uniform publisher coverage than any of the other seven databases. For example, it includes citations for 83% of the core articles published by Sage, 96% of those published by the Gerontological Society of America, and 67% of those published by the US Department of Agriculture. The corresponding figures for SSCI are 69%, 100%, and 0%, respectively. (The standard deviation values in Table 2 show the degree of variability within each database.)

Greater variability can be seen where abstracts and full-text links are concerned. (See Table 3.) Still, there are no major publishers of core articles for which GS provides less than 40% abstract coverage. Google Scholar's full-text coverage varies substantially by publisher, perhaps because GS relies almost entirely on its publishing partners for the provision of full-text links. For instance, there are no links to articles published by the Rural Sociological Society, since the RSS journals are not available at the Society's Web site.

Table 3
Detailed publisher coverage of Google Scholar

	Citations	Abstracts	FT links
Number of core articles for which citations, abstracts, and full-text links are provided			
Sage (35)[a]	29	14	4
Blackwell (34)	34	24	13
Gerontological Society of America (24)	23	23	0
Rural Sociological Society (8)	8	4	0
Springer (7)	7	7	6
Hodder Arnold (5)	5	3	1
Pion (5)	4	3	0
Wiley (5)	5	5	5
Community Development Society (3)	3	3	0
Haworth (3)	2	2	0
Population Association of America (3)	3	3	3
US Department of Agriculture (3)	2	1	0
American Economic Association (2)	2	2	2
Arizona Business (2)	1	0	0
16 Publishers, each with 1 article (16)	16	3	2
Percentage of core articles for which citations, abstracts, and full-text links are provided			
Sage (35)[a]	83	40	11
Blackwell (34)	100	71	38
Gerontological Society of America (24)	96	96	0
Rural Sociological Society (8)	100	50	0
Springer (7)	100	100	86
Hodder Arnold (5)	100	60	20
Pion (5)	80	60	0
Wiley (5)	100	100	100
Community Development Society (3)	100	100	0
Haworth (3)	67	67	0
Population Association of America (3)	100	100	100
US Department of Agriculture (3)	67	33	0
American Economic Association (2)	100	100	100
Arizona Business (2)	50	0	0
16 Publishers, each with 1 article (16)	100	19	13

[a] Numbers in parentheses indicate the number of core articles from each publisher.

Table 4
Date coverage of Google Scholar and seven other databases

	1990 (30)[a]	1991 (15)	1992 (16)	1993 (20)	1994 (8)	1995 (17)	1996 (14)	1997 (13)	1998 (4)	1999 (9)	2000 (9)	Std. dev.
Number of core articles for which citations are provided												
Google Scholar	28	14	13	19	7	17	12	13	4	8	9	–
SSCI	20	14	11	17	4	14	11	8	3	5	6	–
ArticleFirst	16	9	9	11	4	13	9	8	3	6	6	–
Soc. Sci. Abs.	19	12	9	13	4	9	7	8	1	3	1	–
AgeLine	21	6	11	6	3	7	8	2	3	2	4	–
POPLINE	9	6	11	8	4	9	7	7	0	2	4	–
Acad. Srch. Elite	10	2	5	8	3	9	9	6	2	5	3	–
GEOBASE	6	4	7	7	4	9	6	8	2	2	5	–
Percentage of core articles for which citations are provided												
Google Scholar	93	93	81	95	88	100	86	100	100	89	100	7
SSCI	67	93	69	85	50	82	79	62	75	56	67	13
ArticleFirst	53	60	56	55	50	76	64	62	75	67	67	9
Soc. Sci. Abs.	63	80	56	65	50	53	50	62	25	33	11	20
AgeLine	70	40	69	30	38	41	57	15	75	22	44	20
POPLINE	30	40	69	40	50	53	50	54	0	22	44	19
Acad. Srch. Elite	33	13	31	40	38	53	64	46	50	56	33	14
GEOBASE	20	27	44	35	50	53	43	62	50	22	56	14

[a] Numbers in parentheses indicate the number of core articles published each year.

5.3. Date coverage

Several authors have expressed concerns about Google Scholar's date coverage (Burright, 2006; Neuhaus et al., 2006; Tennant, 2005a, 2005b; Vine, 2006), most often criticizing its handling of very recent articles. This analysis provides a broader overview, showing how date coverage varies for core articles published from 1990 to 2000. (See Table 4.) For several of the databases, there is a trend toward better citation coverage of the articles published later in the decade. This is true of Google Scholar as well. For GS, there is a moderate correlation between year of publication and the percentage of core articles for which citations are provided; $r\ (9) = .38$. At the same time, however, Google Scholar displays the least variability in its date coverage—far less than AgeLine, POPLINE, or Social Sciences Abstracts, for instance.

As with publisher coverage, there is greater variability in date coverage where abstracts and full-text links are concerned. (See Table 5.) Curiously, Google Scholar provides abstracts for *fewer* of the core articles published later in the decade. There is a moderate inverse relationship between year of publication and the percentage of core articles for which abstracts are provided; $r\ (9) = .41$. This pattern makes GS unique among the eight databases. The others either provide no abstracts at all (ArticleFirst), provide abstracts for more of the articles published in recent years (Social Sciences Abstracts), or provide abstracts consistently for all 11 years (SSCI, AgeLine, POPLINE, Academic Search Elite, and GEOBASE).

Table 5
Detailed date coverage of Google Scholar

	1990 (30)[a]	1991 (15)	1992 (16)	1993 (20)	1994 (8)	1995 (17)	1996 (14)	1997 (13)	1998 (4)	1999 (4)	2000 (9)
Number of core articles for which citations, abstracts, and full-text links are provided											
Citations	28	14	13	19	7	17	12	13	4	8	9
Abstracts	22	9	11	13	4	10	7	8	2	5	6
FT links	3	4	4	4	0	4	1	4	0	5	7
Percentage of core articles for which citations, abstracts, and full-text links are provided											
Citations	93	93	81	95	88	100	86	100	100	89	100
Abstracts	73	60	69	65	50	59	50	62	50	56	67
FT links	10	27	25	20	0	24	7	31	0	56	78

[a] Numbers in parentheses indicate the number of core articles published each year.

Google Scholar's provision of full-text links is similar to that of the other databases. Full-text links are more likely to be provided for papers published later in the decade; r (9) = .52.

5.4. Sources of citations and abstracts

Google has not revealed the names of its Google Scholar partners, except for those who agreed to participate before the initial beta release of the database (Jacsó, 2004). The sources of Google Scholar's citations, abstracts, and full-text links within the field of later-life migration can be seen in Table 6, however. Thirty-one percent of the core citations were acquired from PubMed. An approximately equal number of citations—those marked "[CITATION]" on the Google Scholar results pages—were extracted from the reference lists of other documents. (Citations extracted from other documents include only basic information: author, article title, journal title, and year of publication.) Altogether, 36% of the GS citations were acquired from publicly available databases (PubMed and ERIC), 17% from commercial publishers, 8% from scholarly societies, and 8% from JSTOR and Ingenta.

For this set of core articles, the Google Scholar partners provides an abstract for nearly every citation. (All but one of the citations without abstracts were extracted from the reference lists of other documents.) While many scholarly societies and public agencies provide online access to their full-text document collections, this is not reflected in the results for this particular set of articles. As Table 6 shows, full-text links to core articles

Table 6
Sources of citations, abstracts, and full-text links for core articles in Google Scholar

	Number of citations from source	Number with abstracts	Number with full-text links	Percentage of citations from source	Percentage with abstracts	Percentage with full-text links
[CITATION][a]	46	0	0	32	0	0
PubMed	44	44	0	31	100	0
GSA[b]	10	10	0	7	100	0
JSTOR	10	10	10	7	100	100
Blackwell	9	8	9	6	89	100
ERIC	7	7	0	5	100	0
Springer	7	7	6	5	100	86
Wiley	5	5	5	3	100	100
Sage	4	4	4	3	100	100
IASSIST[b]	1	1	1	1	100	100
Ingenta	1	1	1	1	100	100

[a] These citations, marked "[CITATION]" on the GS search results pages, were extracted from the reference lists of other documents. They include only basic information: author, article title, journal title, and year of publication.
[b] GSA is the Gerontological Society of America. IASSIST is the International Association for Social Science Information Service and Technology.

Table 7
Overlaps in citation coverage between Google Scholar and seven other databases

	SSCI	Article-First	Soc. Sci. Abs.	Age-line	POP-LINE	Acad. Srch. Elite	GEO-BASE
Number of core articles with records (citations) in GS	144	144	144	144	144	144	144
Number of core articles with records (citations) in database	113	94	86	73	67	62	60
Number of records (citations) found in both GS and database	108	89	83	64	65	62	59
Percentage of database citations that are also included in GS	96	95	97	88	97	100	98
Percentage of GS citations that are also included in database	75	62	58	44	45	43	41

are provided by all the commercial journal publishers but not by the Gerontological Society of America, PubMed, or ERIC.

5.5. Overlaps in citation coverage

The breadth of Google Scholar's coverage can be seen in Table 7, which shows that most of the core articles indexed by the other seven databases can be found in Google Scholar as well. In fact, with one exception (AgeLine), GS includes citations to at least 95% of the core articles covered by each of the other databases.

Table 7 also shows how well the other seven databases index the core articles found in Google Scholar. Perhaps not surprisingly, those that most effectively cover the literature of later-life migration—SSCI, ArticleFirst, and Social Sciences Abstracts—are also most similar to GS in their coverage of this subject area. However, no database indexes more than 75% of the core articles included in Google Scholar, and GS provides records for eight core articles that are not indexed by any of the other seven databases. Two of the eight appeared in popular periodicals, two appeared in a government-sponsored journal (*Rural Development Perspectives*), and two appeared in the *Journal of the Community Development Society*.

6. Conclusion

Because Google Scholar's search, retrieval, and record management mechanisms are relatively unsophisticated, the current version of GS is unlikely to replace conventional social science databases for serious scholarly work. Google Scholar is nonetheless valuable due to its comprehensive coverage. Within the field of later-life migration, GS indexes 27% more core articles than the second-ranked database (SSCI) and 2.4 times as many core articles as the lowest-ranked database (GEOBASE). Moreover, it covers 88–100% of the core articles found in each of the other seven databases.

Despite Google Scholar's reliance on cooperative arrangements with particular publishers and vendors, there is no evidence of publisher bias in its citation coverage. Unlike SSCI and Social Sciences Abstracts, for instance, GS provides roughly equal coverage of every publisher's articles. Google Scholar is also remarkably consistent in its date coverage, providing the most comprehensive indexing of core articles from the early 1990s despite the fact that it was released just two years ago.

The results of this analysis are inconsistent with previous reports in at least two respects. Although several authors have criticized Google Scholar's inclusion of duplicate records and records representing the less authoritative versions of scholarly works (Bar-Ilan, 2006; Jacsó, 2004; Tenopir, 2005), neither problem was encountered more than a few times in the course of this investigation. Only a few core articles were represented more than once within the GS search results, and only two were cited in multiple manifestations (as both journal articles and preliminary reports). At the same time, these results suggest that the popular media (Wray, 2004, among others) may have overstated the importance of Google Scholar as a provider of free access to scholarly papers. While GS does provide links to many full-text documents, only one core article was available free of charge. Admittedly, the situation may be different for papers published since 2000.

As noted earlier, many of Google Scholar's deficiencies involve its search mechanism rather than its content. However, this evaluation suggests two areas in which the database content might be improved. First, Google Scholar provides relatively few abstracts, especially in comparison with other free databases such as AgeLine and POPLINE. (GS ranks sixth of the eight databases in the proportion of core article records for which abstracts are provided.) Second, nearly a third of the core article citations lack volume numbers, issue numbers, and page numbers. Both these problems can be traced to the inclusion of records extracted from the reference lists of other documents—virtually the only GS records for which abstracts are missing and citations are incomplete.

This study confirms that a list of relevant items, compiled in advance, can be used to evaluate database coverage in a meaningful way. At the same time, the use of only 155 core articles—all published prior to 2001—precludes an assessment of Google Scholar's coverage of the most recent literature. Moreover, the coverage (content) of each database is just one component of its effectiveness. Because few, if any, subject searches can be expected to retrieve the entire set of relevant documents, the ultimate value of GS can be determined only through an evaluation of its search mechanisms as well as its content.

The importance of Google Scholar within any library or research organization will depend on the characteristics of the institution as well as the characteristics of the database. Librarians attempting to place GS within the broader range of library collections and services may want to consider a number of questions:

- Is the library's primary mission to provide information to clients, or to educate students?
- Does the library offer a full range of disciplinary databases? Do multidisciplinary databases such as Google Scholar supplement them or replace them?
- How important is the development of patrons' database searching skills, either for use in later research or as a means of encouraging critical thinking and precision in academic work?
- What proportion of the items indexed in GS can the library provide, either immediately or through interlibrary loan? Is the presence of records for inaccessible items an advantage or a problem? (Do patrons have a strong interest in identifying all relevant research, or will they be discouraged by the presence of many records for items that are not immediately accessible?)

Google Scholar provides excellent coverage of at least one social science field, in both absolute and comparative terms. The results presented here suggest that the unconventional methods used to build the GS database have not adversely affected its coverage of the scholarly literature.

Acknowledgements

I am grateful for the comments of Esther Isabelle Wilder and three anonymous referees.

References

Bar-Ilan, J. (2006). An ego-centric citation analysis of the works of Michael O. Rabin based on multiple citation indexes. *Information Processing and Management, 42*(6), 1553–1566.

Bauer, K., & Bakkalbasi, N. (2005). An examination of citation counts in a new scholarly communication environment. *D-Lib Magazine, 11* (9). http://www.dlib.org/dlib/september05/bauer/09bauer.html.

Burright, M. (2006). Google Scholar: science & technology. *Issues in Science & Technology Librarianship, 45.* http://www.istl.org/06-winter/databases2.html.

Findlay, A. M. (1991). Population geography. *Progress in Human Geography, 15*(1), 64–72.

Findlay, A. M. (1992). Population geography. *Progress in Human Geography, 16*(1), 88–97.

Findlay, A. M. (1993). Population geography: disorder, death and future directions. *Progress in Human Geography, 17*(1), 73–83.

Gardner, S., & Eng, S. (2005). Gaga over Google? Scholar in the social sciences. *Library Hi Tech News, 22*(8), 42–45.

Giustini, D., & Barsky, E. (2005). A look at Google Scholar, PubMed, and Scirus: comparisons and recommendations. *Journal of the Canadian Health Libraries Association, 26*(3), 85–89. http://pubs.nrc-cnrc.gc.ca/jchla/jchla26/c05-030.pdf.

Google (2005). About Google Scholar. http://scholar.google.com/intl/en/scholar/about.html.

Google (2006). Google Scholar. http://scholar.google.com.

Gorman, G. E. (2006). Giving way to Google. *Online Information Review, 30*(2), 97–99.

Griffith, B. C., White, H. D., Drott, M. C., & Saye, J. D. (1986). Tests of methods for evaluating bibliographic databases: an analysis of the National Library of Medicine's handling of literatures in the medical behavioral sciences. *Journal of the American Society for Information Science, 37*(4), 261–270.

Hamaker, C., & Spry, B. (2005). Key issue: Google Scholar. *Serials, 18*(1), 70–72. http://serials.uksg.org/openurl.asp?genre=article&issn=0953-0460&volume=18&issue=1&spage=70.

Jacsó, P. (2004). Péter's picks & pans: CiteBaseSearch, Institute of Physics Archive, and Google's Index to Scholarly Archive. *Online, 28*(5), 57–60.

Jacsó, P. (2005a). As we may search: comparison of major features of the Web of Science, Scopus, and Google Scholar citation-based and citation-enhanced databases. *Current Science, 89*(9), 1537–1547. http://www.ias.ac.in/currsci/nov102005/1537.pdf.

Jacsó, P. (2005b). Comparison and analysis of the citedness scores in Web of Science and Google Scholar. In E. A. Fox, E. J. Neuhold, P. Premsmit, & V. Wuwongse (Eds.), *Digital libraries: implementing strategies and sharing experiences*. *Lecture Notes in Computer Science* (3815, pp. 360–369). Berlin: Springer-Verlag.

Jacsó, P. (2005c). Evaluation of citation enhanced scholarly databases. INFOPRO 2005 Keynote Address, Tokyo, 10 November 2005. http://tokyo-keynote.notlong.com/.

Jacsó, P. (2005d). Google Scholar: the pros and the cons. *Online Information Review, 29*(2), 208–214.

Jacsó, P. (2005e). Visualizing overlap and rank differences among Web-wide search engines: some free tools and services. *Online Information Review, 29*(5), 554–560.

Jacsó, P. (2006). Deflated, inflated and phantom citation counts. *Online Information Review, 30*(3), 297–309.

Jones, D.Y. (2005). Biology article retrieval from various databases: making good choices with limited resources. *Issues in Science and Technology Librarianship, 44*. http://www.istl.org/05-fall/refereed.html.
Mullen, L. B., & Hartman, K. A. (2006). Google Scholar and the library Web site: the early response by ARL libraries. *College and Research Libraries, 67*(2), 106–122.
Myhill, M. (2005). Google Scholar. *Charleston Advisor, 6* (4), 49–52. http://www.charlestonco.com/review.cfm?id=225.
Nash, A. (1994a). Population geography. *Progress in Human Geography, 18*(1), 84–91.
Nash, A. (1994b). Population geography. *Progress in Human Geography, 18*(3), 385–395.
Nash, A. (1996). Population geography. *Progress in Human Geography, 20*(2), 203–214.
Neuhaus, C., Neuhaus, E., Asher, A., & Wrede, C. (2006). The depth and breadth of Google Scholar: an empirical study. *Portal: Libraries and the Academy, 6*(2), 127–141.
Notess, G. R. (2005). Scholarly Web searching: Google Scholar and Scirus. *Online, 29*(4), 39–41.
Ogden, P. E. (1998). Population geography. *Progress in Human Geography, 22*(1), 105–114.
Ogden, P. E. (2000). Weaving demography into society, economy and culture: progress and prospect in population geography. *Progress in Human Geography, 24*(4), 627–640.
Ojala, M. (2005). Scholarly mistakes. *Online, 29*(3), 26.
O'Lear, S. (1997). Migration and the environment: a review of recent literature. *Social Science Quarterly, 78*(2), 606–618.
O'Leary, M. (2005). Google Scholar: what's in it for you? *Information Today, 22*(7), 35–39.
Peek, R. (2005). A Googly new year. *Information Today, 22*(1), 17–18.
Tennant, R. (2005a). Google, the naked emperor. *Library Journal, 130*(13), 29.
Tennant, R. (2005b). Is metasearching dead? *Library Journal, 130*(12), 28.
Tenopir, C. (1982). Evaluation of database coverage: a comparison of two methodologies. *Online Review, 6*(5), 423–441.
Tenopir, C. (2005). Google in the academic library. *Library Journal, 130*(2), 32.
Vine, R., (2006). Google Scholar. *Journal of the Medical Library Association, 94*(1), 97–99. http://www.pubmedcentral.nih.gov/picrender.fcgi?artid=1324783&blobtype=pdf.
Walters, W. H. (2002). Later-life migration in the United States: a review of recent research. *Journal of Planning Literature, 17*(1), 37–66.
Walters, W. H., & Wilder, E. I. (2003). Bibliographic index coverage of a multidisciplinary field. *Journal of the American Society for Information Science and Technology, 54*(14), 1305–1312.
Warnes, A. M. (1990). Geographical questions in gerontology: needed directions for research. *Progress in Human Geography, 14*(1), 24–55.
Wleklinski, J. M. (2005). Studying Google Scholar: wall to wall coverage? *Online, 29*(3), 22–26.
Wray, R., (2004) November 22. Google puts new slant on scholarship. *The Guardian* (London), Final Edition, 7.

William H. Walters (Ph.D., Brown University) is an Assistant Professor of Librarianship at Millersville University. His research has appeared in nearly 20 journals including those of the Royal Geographical Society, the Gerontological Society of America, the Regional Science Association International, the British Cartographic Society, and the American Society for Information Science and Technology.