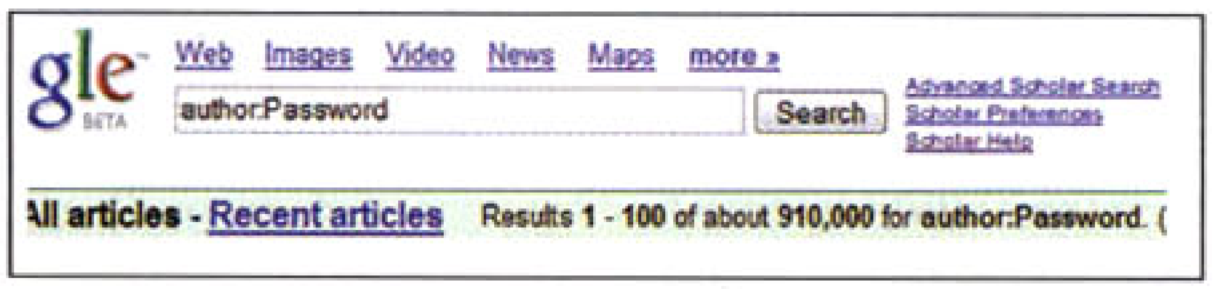
GOOGLE SCHOLAR'S GHOST AUTHORS

By Peter Jacso

The popular tool can't be used to analyze the publishing performance and impact of researchers

Geoffrey Nunberg's August 31 essay in *The Chronicle of Higher Education* criticizing Google's Book Search (GBS), which he subtitled "A Disaster for Scholars," emphasized that disturbing errors are endemic. He well recognizes that for mainstream "googling" purposes, "we don't really care about metadata...provided by a library catalog." In perhaps his most discouraging point, linguistics professor Nunberg notes that the Google team blames libraries and publishers for bad data.



author:Password

Results 1 - 100 of about 910,000 for author:Password. (

CENSUS Google Scholar claimed a million authors named "Password"

All these rhyme perfectly with my experience working with another of the search giant's data-crunching products, Google Scholar (GS). With GS (scholar.google.com), however, I blame mostly the developers. They decided—very unwisely—not to use the good metadata generously offered to them by scholarly publishers and indexing/abstracting services but instead chose to figure them out through ostensibly smart crawler and parser programs.

Thus research faculty and academic/special libraries dealing with GS face their own metadata disaster, one with dire consequences in evaluating the scholarly publishing productivity and impact of researchers, institutions, journals, and even countries. Millions of records have erroneous metadata, as well as inflated publication and citation counts, creating "ghost authors," like "Password" (pictured here), and "lost authors."

The free Google Scholar has become the most convenient resource to find a few good scholarly papers—often in free full-text format—on even the most esoteric topics. For topical keyword searches, GS is most valuable. But it cannot be used to analyze the publishing performance and impact of researchers.

Peter Jacso is Professor and Chair of the Library and Information Science Program in the Department of Information and Computer Sciences at the University of Hawai'i at Manoa

Citation problems

Google's algorithms create phantom authors for millions of papers. They derive false names from options listed on the search menu, such as P Login (for Please Login).

Very often, the real authors are relegated to ghost authors deprived of their authorship along with publication and citation counts. In the scholarly world, this is critical, as the mantra "publish or perish" changes to "publish, get cited, or perish."

Compounding the problem, the inflated publication and citation counts produced by GS will embarrass those who take the reported numbers at face value, as they discover that many of the publications, randomly scattered throughout the detailed result lists, are just variant formats of the same paper, and the citations are mismatched.

Skewed data

Google Scholar's publication/citation counts and metadata for bibliometric and scientometric evaluations too often resemble Bernie Madoff's profit numbers. Just as investors preferred the nonexistent reality described by Madoff's tally, users may like the publication/citation counts reported by GS and the many inflated indicators derived from them. The numbers in GS are inflated for many reasons. First and foremost, GS lumps together the number of master records (created for actual publications) and the number of citation records (distinguished by the prefix: [citation]) when reporting the total hits for an author name search.

By contrast, fee-based Web of Science and Scopus have lower article and citation counts and scientometric indicators; they have a far more selectively defined source base with fewer journals from which to gather publication and citations data. In addition, they count only the master records for the authors' publication count (as they should) and keep the stray and orphan citations in a separate file. These stray and orphan citations in Web of Science and Scopus are normalized manually only in exceptional cases (because it is very tedious).

For the full version of this article from *LJ Academic Newswire*, with more screenshots and links, see bit.ly/JacsoGS

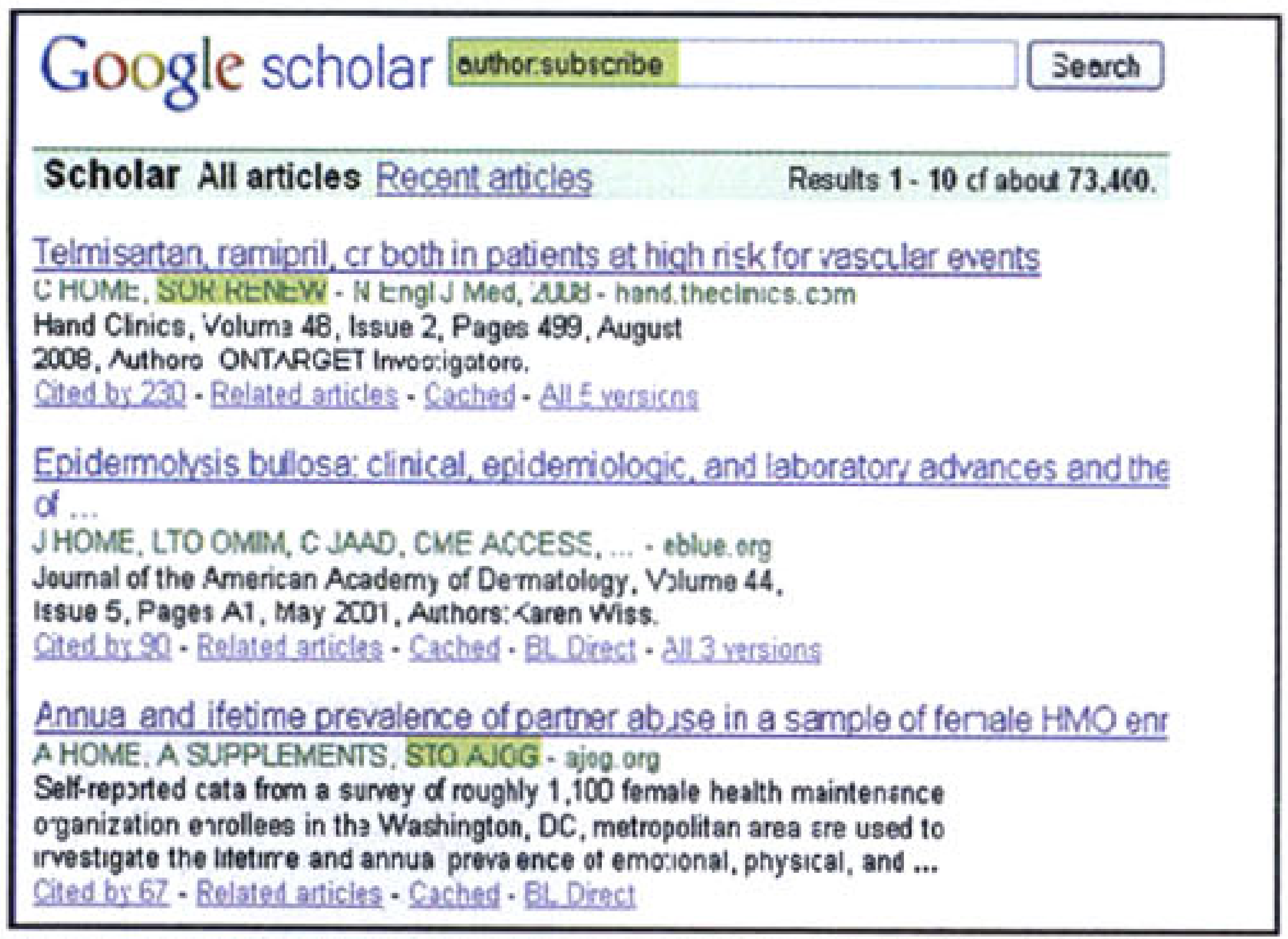
Compounding errors

Unfortunately, the bad metadata has a long reach. These numbers are taken at face value by the free utilities such as the Google Scholar Citation Count gadget developed by Jan Feyereisl and the sophisticated and pretty Publish or Perish (PoP) software produced by Tarma Software.

Such utilities turn many people into neophyte citation analysts who don't see, don't want to see, or assertively deny the metadata mess in GS and produce ranking lists of researchers and journals based on both metadata and publication and citation counts reported by GS.

Ghosts in the machine

Since about 10.2 million records from GBS are incorporated now into GS, the metadata disaster likely will continue unabated. It is bad enough to have so many records with erroneous publication years, titles, authors, and journal names. It becomes much worse with millions of fabricated ghost author names.



Results 1 - 10 cf about 73,400.

Telmisartan, ramipril, cr both in patients at high risk for vasculer events
C HOME, SOR RENEW - N Engl J Med, 2008 - hand.theclinics.com
Hand Clinics, Volume 48, Issue 2, Pages 499, August 2008, Authors ONTARGET Investigators.
Cited by 230 - Related articles - Cached - All 5 versions

Epidermolysis bullosa: clinical, epidemiologic, and laboratory advances and the of ...
J HOME, LTO OMIM, C JAAD, CME ACCESS, ... - eblue.org
Journal of the American Academy of Dermatology, Volume 44, Issue 5, Pages A1, May 2001, Authors: Karen Wiss.
Cited by 90 - Related articles - Cached - BL Direct - All 3 versions

Annua and ifetime prevalence of partner abuse in a sample of female HMO enr
A HOME, A SUPPLEMENTS, STO AJOG - ajog.org
Self-reported cata from a survey of roughly 1,100 female health maintenance organization enrollees in the Washington, DC, metropolitan area are used to investigate the lifetime and annual prevalence of emotional, physical, and ...
Cited by 67 - Related articles - Cached - BL Direct

WHAT'S IN A NAME? Subscribe is a common author name in Google Scholar. It may appear in masked form under the initials STO or SOR if the menu option is Subscribe To or Subscribe or Renew

False names are created from options on the search menu, such as P Options (for Payment Options), from parts of the author affiliation (CA San Diego, C Ltd, M View for Mountain View), from Table of Contents pages on publishers' web sites, and from section headings of articles (B Methods, D Definitions, G Assessment, H Variables, I Evaluation).

Subscribe (73,400 results) seems to be a common author name. It may not be easy to spot just by browsing the results list, however, because it may appear under the initials STO or SOR if the menu option is Subscribe To or Subscribe or Renew (see graphic).

The parser knows no hurdles and fabricates a single initial or several initials from the letter or Roman numerals preceding the section titles—I Background, V Findings, X Conclusions—and from the first word of the menu options of the homepage P Login (from Please Login), N Subscriber (from New Subscriber), A Registered (from Already Registered), SD Access (from Science Direct Access).

GS ignores correct publication years, fancying page numbers, volume numbers, parts of document codes, ZIP codes, and street addresses of author affiliations as publication years instead.

Lost authors abound

These errors could be considered relatively harmless if they did not affect the contributions of genuine, real-live scholars. But the biggest problem is when the mess replaces real scholars with ghost authors, leaving the former as lost authors, robbed of their authorship and citations.

How does the parser work? In one case, it took from the table of contents page the title of one paper, fabricated the first initials and the last name of the first author from the subtitle of another paper and the second author's name from the title of a third paper. This is a joke and a very bad one.

This happens very often in records for papers in *The Lancet*, but this type of error is endemic. It may be that the data harvested from *The Lancet* is on the route of the most undertrained crawler/parser puppy GS has unleashed.

Attributing errors

Certainly, the entire database isn't rotten—just a few million records. That may be a relatively small percentage—Google won't reveal the total number of records, and these are just my few forensic search test queries—but there's cause for worry.

In the case of GBS, Google relied on its collective Pavlovian reflex to blame the publishers and libraries (librarians, catalogers, indexers) for the wrong metadata. In the case of Google Scholar, these same Googlish arguments will not fly, because practically all the scholarly publishers gave Google—hats in hand—their digital archive with metadata. The idea was to have Google index it and drive traffic to the publishers' sites.

Yes, GS has fixed fairly quickly some of the major errors I have used to demonstrate its illiteracy and innumeracy but has so far left millions of others untouched. I am happy that I no longer see many of my most disliked phantom authors fabricated by Google Scholar, such as members of the Password family once credited with authoring 910,000 papers.

How did we get here?

It must have taken some time to create such an imbecile parser. In the early days, the GS developers decided not to use the metadata readily available from most of the scholarly publishers. This is obvious from the highly improved, intelligent (free) Scirus system that has made smart use of the publishers' metadata after its first bad steps that I criticized upon its debut.

The press and the public were so enamored of anything with the word Google in it that GS developers apparently believed they could create a parser to identify the metadata better than human indexers. Not all of the indexing/abstracting services are perfect and consistent, but their errors are dwarfed by the types and volume of those in GS.

Just as with GBS, commercial passion is the deciding factor for Google. So I am far less optimistic than Nunberg about Google's pledges to improve the metadata "train wreck" (to borrow his term).

The parsers have not improved significantly in the past five years despite much criticism. GS developers corrected some errors that got negative publicity, but these were Band-Aids, where brain surgery and extensive parser training is required. ■