NBER WORKING PAPER SERIES

PUBLICATIONHARVESTER: AN OPEN-SOURCE SOFTWARE TOOL FOR POLICY RESEARCH

Pierre Azoulay
Andrew Stellman
Joshua Graff Zivin

Working Paper 12039
http://www.nber.org/papers/w12039

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
February 2006

Send all correspondence to pa2009@columbia.edu or stellman@stellman-greene.com. Part of the work was performed while the first author was an Alfred P. Sloan Industry Studies Fellow. We gratefully acknowledge the financial support of the Merck Foundation through the Columbia-Stanford Consortium on Medical Innovation. The usual disclaimer applies. The views expressed herein are those of the author(s) and do not necessarily reflect the views of the National Bureau of Economic Research.

PublicationHarvester: An Open-Source Software Tool for Science Policy Research
Pierre Azoulay, Andrew Stellman, Joshua Graff Zivin
NBER Working Paper No. 12039
February 2006
JEL No. O32

ABSTRACT

We present PublicationHarvester, an open-source software tool for gathering publication information on individual life scientists. The software interfaces with MEDLINE, and allows the end-user to specify up to four MEDLINE-formatted names for each researcher. Using these names along with a user-specified search query, PublicationHarvester generates yearly publication counts, optionally weighted by Journal Impact Factors. These counts are further broken-down by order on the authorship list (first, last, second, next-to-last, middle) and by publication type (clinical trials, regular journal articles, reviews, letters/editorials, etc.) The software also generates a keywords report at the scientist-year level, using the Medical Subject Headings (MeSH) assigned by the National Library of Medicine to each publication indexed by Medline. The software, source code, and user manual can be downloaded at http://www.stellman-greene.com/PublicationHarvester/

Pierre Azoulay
Columbia University
Graduate School of Business
3022 Broadway, Uris Hall 704
New York, NY 10027
and NBER
pa2009@columbia.edu

Joshua Graff Zivin
Department of Health Policy and Management
Columbia University
600 West 168th Street, Room 608
New York, NY 10032
and NBER
jz126@columbia.edu

Andrew Stellman
Stellman & Greene Consulting LLC
117 St. John's Place
Brooklyn, NY 11217
astellman@stellman-greene.com

1 Introduction

Many innovation and science policy scholars make heavy use of publication data. However, researchers face a number of practical constraints when gathering such data, especially when the unit of analysis is the individual scientist. *PublicationHarvester* is an open-source software tool that automates the process of gathering publication information for individual life scientists. It is fast, simple to use, and reliable.

2 Data Sources

PublicationHarvester searches `MEDLINE` and `OLDMEDLINE`, two bibliographic databases maintained by the U.S. National Library of Medicine that are searchable on the web at no cost.[1] They contain over 14 million citations from 4,800 journals published in the United States and more than 70 other countries from 1950 to the present. The subject scope of these databases is biomedicine and health, broadly defined to encompass those areas of the life sciences, behavioral sciences, chemical sciences, and bioengineering that inform research in health-related fields. As a result, *PublicationHarvester* will be of limited usefulness to innovation scholars who use other scientific fields as settings for their investigations.

3 Publication Data at the Individual Level

Researchers face two distinct issues when attempting to generate publication counts for individual scientists. The first relates to the uniqueness of names: common names make it difficult to distinguish between scientists, and scientists with relatively rare names sometimes are inconsistent in their use of publication names. The second is quality adjustment. *PublicationHarvester* allows the end-user to specify a search query for each scientist so that his/her publication output is comprehensive and can be distinguished from that of other scientists. With this work accomplished, the software quality-adjusts publication counts by

[1] `http://www.pubmed.gov/`

using information about the scientific impact of individual journals, publication types, and order of authorship.

3.1 Dealing with frequent or inconsistent names

Namesakes and popular names. The two most comprehensive publication databases, `MEDLINE` and the Scientific Citation Index (SCI) do not assign unique identifiers to the authors of the publications they index. They identify authors simply by their last name, up to two initials (three in the SCI), and an optional suffix (in `MEDLINE`). This makes it difficult to unambiguously assign publication output to individual scientists, especially when their last name is relatively common.

Researchers can deal with this problem in three different ways. The first solution is to ignore it altogether. This may be less problematic that commonly thought, since name popularity should be, to a first order of approximation, orthogonal to other determinants of scientific productivity.[2]

An easy-to-implement alternative is to make use of name frequencies as weights to account for the relatively greater uncertainty surrounding publication counts for scientists with popular names. Of course, this assumes that researchers have at their disposal an auxiliary data source from which these frequencies can be computed. Azoulay and Graff Zivin (2005) use the faculty roster of the American Association of Medical Colleges to compute frequency weights for every combination of last name and two initials. For example, if there are 9 researchers in the faculty roster whose last name is "Jones", whose first name begins by "J", and whose middle initial is "E," then in any regression with publication counts as a dependent variable, the observations corresponding to James E. Jones and J. Edwin Jones could be weighted by $\frac{1}{9} \simeq 0.111$.[3]

[2] Assuming that the publication counts are dependent variables in regression equations, measurement error of the classical kind results in inflated standard errors, but does not affect the consistency of the estimates (Wooldridge 2002: 74.)

[3] Of course, this approach assumes all scientists are equally productive, which will be a more acceptable assumption for some research questions than others. More sophisticated approaches to distinguish between authors are feasible, for example by using co-author inclusion recursively (Wooding et al. 2005). However, these approaches are computationally demanding when dealing with a large number of scientists.

With relatively small samples, a more labor-intensive (but more precise) approach is to narrow the list of publications using boolean operators, scientific keywords, and institutional affiliations. MEDLINE supports elaborate search queries of this type, and our software allows the user to customize search queries for individual researchers. As an example, one can examine the publications of Keith W. Miller, the Mallinckrodt Professor of Pharmacology in Anesthesia at Harvard Medical School. Miller is a relatively frequent name in the United States (with 702 researchers with an identical patronym in the AAMC faculty roster); the combination `"miller kw"` is common to 3 researchers in the same database. A simple search query for `"miller kw"[au]` returns 203 publications at the time of this writing. However, a more refined query, based on Professor Miller's CV found on the web would return only 145 publications.[4]

Inconsistent publication names. While the preceding section discussed the possibility of gathering too many publications for a researcher, the opposite danger, that of recording too few publications, also looms large, since scientists are often inconsistent in the choice of names they choose to publish under. By far the most common source of error is the haphazard use of a middle initial. Other errors stem from inconsistent use of suffixes (Jr., Sr., 2nd, etc.), multiple patronyms due to changes in spousal status (e.g., `"graff-zivin j"` vs. `"zivin jg"`), or, more rarely, from inconsistent use of a first initial (e.g., M. Judah Folkman's publications can be found under both `"folkman j"` and `"folkman mj"`).

To deal with this problem, *PublicationHarvester* gives the end-user the option to choose up to four MEDLINE-formatted names under which publications can be found for a given researcher. For example, Louis J. Tobian, Jr. publishes under `"tobian l"`, `"tobian l jr"`, and `"tobian lj"`, and all three names need to be provided as input to generate a complete publication listing. Furthermore, even though Tobian is a relatively rare name, the search query needs to be modified to account for these name variations, as in `("tobian l"[au] OR "tobian lj"[au])`.[5]

[4] `("miller kw"[au] AND (biomembr* OR bilaye* OR submar* OR decompres* OR permeab* OR bubbl* OR barbitur* OR ligands OR alchoh* OR acetylcholin* OR chromatograp* OR cimetidine OR pharmacot* OR pharmacol* OR anesthes* OR anesthet* OR pharmacol*))`

[5] Other examples are provided in the sample input file that can be downloaded from the web site.

3.2 Quality-adjustment: Alternatives to citation-weighting

Science policy researchers have long recognized that scientific output tends to be very heterogeneous in quality, with a small number of researchers responsible for a disproportionate fraction of total output in any given field (Lotka, 1926). Even for a given scientist, a small number of publications will typically account for the bulk of citations to that researcher's publication output. As a result, scholars in the fields of the economics and sociology of science have often viewed citation-weighting at the article level as the "gold standard" against which other quality-adjustment methods need to be evaluated. Unfortunately, article-level citation data is expensive to gather when dealing with a very large number of articles. This has led researchers to look for cheaper, yet sensible alternatives.[6] This is the direction adopted here. The software can capture two dimensions of article quality: journal prominence and publication type.

Journal Impact Factors. We make use of the Journal Citation Reports, published yearly by the Institute for Scientific Information. ISI ranks journals by impact factor (JIF) in different scientific fields. The impact factor is a measure of the frequency with which the "average article" in a journal has been cited in a particular year. The software weights each article published by the scientists in our sample by the corresponding journal's average JIF, and computes quality-weighted publication counts in this way.[7]

Publication Types. In many scientific fields, researchers publish many articles a year, but only a fraction of those correspond to genuinely new pieces of scientific knowledge. Others are reviews, invited comments, errata, editorials, etc. *PublicationHarvester* gives the end-user the option to break down publication output into five broad different publication types. In so doing, the software makes use of the publication type field attached to every article indexed by Medline. In the file `sample-pubtypes.csv`, we have aggregated the 51 Medline

[6] In some cases, there is no workable alternative to the use of article-level citation data, in particular when the timing of the stream of citations is at issue, or when information about "who cites whom" is essential to the study (e.g., Murray and Stern 2005).

[7] Basically a ratio between citations and recent citable items published, JIFs suffer from built-in biases: they tend to discount the advantage of large journals over small ones, of frequently-issued journals over less frequently-issued ones, and of older journals over newer ones (Garfield 2006).

publication types into five "logical" broad groups, but other users could without difficulty elect a different aggregation scheme.[8]

3.3 Using order of authorship information

In the natural and physical sciences, a robust social norm systematically assigns last authorship to the principal investigator, first authorship to the junior author who was responsible for the actual conduct of the investigation, and apportions the remaining credit to authors in the middle of the authorship list, generally as a decreasing function of the distance from the extremities (Riesenberg and Lundberg 1990). In the portfolio of any given scientist, the publications in which s/he appears in first or last position will therefore be more important than those where s/he appears in the middle, even holding journal prominence constant. The software breaks down the publication counts according to the first, last, and middle authorship positions. Whenever the number of authors is equal to or greater than five, the software tracks separately the publications in which the author appears in second and next-to-last position. These authorship positions have gained importance over time as the average number of authors per paper grew from 3.37 in 1975 to 6.89 in 2005.[9]

4 Measuring the Direction of Scientific of Research through Keywords

Increasingly, science policy scholars are interested in studying how specific behaviors or policy interventions influence the *direction* of scientific change. For example, Azoulay et al. (2006) use title keywords to compute an index of "patentatibility" for individual researchers. Scientific keywords are also useful when trying to assess how far or close apart two researchers are in "scientific space" (Jaffe 1986). Our software lowers the costs of harvesting scientific

[8]The first group of publications corresponds to clinical research, broadly construed; the second group corresponds to review articles; the third group corresponds to traditional journal articles; and the fourth group comprises letters, comments, and editorials. The last group corresponds to publication types we wish to ignore, such as biographical articles, interviews, festschrifts, etc. This is done simply by assigning those `MEDLINE` publication types an identifier equal to 0.

[9]Authors' tabulations based on an analysis of 721,223 publications by 4,765 "superstar" scientists in the life sciences.

keywords by making use of Medline's MeSH descriptors. These keywords are arranged in both an alphabetic and a hierarchical structure. At the most general level of the hierarchical structure are very broad headings such as `Anatomy` or `Mental Disorders`. More specific headings are found at lower levels of the eleven-level hierarchy, such as `Ankle` and `Conduct Disorder`. There are 22,997 descriptors in the current version of MeSH. Independently of publication counts, the software generate a MeSH report that aggregates the keywords up to the scientist-year level. For example, in the case of Judah Folkman, the software returns 21 publications in 1998. The most frequent keywords for this researcher that year are `Vascular Endothelial Growth Factors`, `Angiostatins`, and `Transfection`. This panel of MeSH keywords is available to researchers for post-processing according to the end-user's idiosyncratic needs.

5 Limitations and Conclusions

PublicationHarvester is not without a set of important limitations. First, it interfaces with `MEDLINE`, so it cannot gather publication and keyword data for researchers outside of the life sciences. Second, it is not useful for organizational-level research, since `MEDLINE` does not list affiliation information.[10] For the same reason, it cannot use article-level citation weights to adjust publication counts for quality. Finally, it puts the onus on the end-user to specify namesakes and queries for individual scientists before harvesting their data, and is therefore better suited to relatively small samples.

Astute readers might wonder why we chose not to query the Science Citation Index, since it contains information about both citations and affiliations. The answer is a practical one. Public access to SCI over the internet is slow, and getting local access is often prohibitively expensive. In contrast, `MEDLINE` access is both fast and free.

[10]More precisely, `MEDLINE` only lists the first author's affiliation (from 1988 onwards). Although the SCI lists affiliation information for all authors from 1973 onwards, it does not link individual authors with their institutions. This seriously limits the usefulness of the SCI affiliation data when conducting research at the individual level of analysis.

Fortunately for the science policy community, other data sources, such as the AAMC Faculty Roster or the NIH Compound Grant Application File (CGAF) contain high-quality affiliation data. Whenever researchers need to track individual scientists across time and space, and simultaneously want to measure their publication output, a cost/benefit analysis would probably favor the use of these administrative data sources in combination with *PublicationHarvester*. Furthermore, for garden-variety quality adjustment, weighting publications by Journal Impact Factor constitutes an easy-to-implement alternative to more precise — but much more expensive — article-level citation weights.

From a computational point of view, our approach also has a number of advantages. The software is extremely fast — it harvested 721,223 publications for 4,765 academics in exactly ten hours. It uses the free relational database `MySQL`[11] to store the harvested data. And it is made freely available under the terms of the GNU General Public License. It is our hope that *PublicationHarvester* will be a valuable tool for other researchers in the innovation and science policy community.

References

Azoulay, P., Graff Zivin, J., 2005. "Peer Effects in the Workplace: Evidence from Professional Transitions for the Superstars of Medicine." Working Paper, Columbia University.

Azoulay, P., Ding, W., Stuart, T., 2006. "The Impact of Academic Patenting on the Rate, Quality, and Direction of (Public) Research Output." NBER Working Paper #11917.

Garfield, E., 2006. "The History and Meaning of the Journal Impact Factor." JAMA 295, 90-93.

Jaffe A.B., 1986. "Technological Opportunity and Spillovers of R&D: Evidences from Firms' Patents, Profits and Market Value." American Economic Review 76, 984-1001.

Murray, F., Stern S., 2005. "Do Formal Intellectual Property Rights Hinder the Free Flow of Scientific Knowledge? An Empirical Test of the Anti-Commons Hypothesis." NBER Working Paper #11465.

Lotka, A.J., 1926. "The Frequency Distribution of Scientific Productivity." Journal of the Washington Academy of Sciences 16, 317-323.

[11] `http://www.mysql.com/`

Riesenberg D., Lundberg G.D., 1990. "The Order of Authorship: Who's on First?" JAMA 264, 1857.

Wooding, S., Wilcox-Jay, K., Lewison, G., Grant, J., 2006. "Co-author Inclusion: A Novel Recursive Algorithmic Method for Dealing with Homonyms in Bibliometric Analysis." Scientometrics 66, 11-21.

Wooldridge, J.M., 2002. Econometric Analysis of Cross Section and Panel Data. The MIT Press, Cambridge, MA.