Sources for Scientific Frustrations: Productivity and Citation Data

Krešimir Zauder
Institute for Social Research in Zagreb
Amruševa 11/II, 10000 Zagreb, Croatia
kresimir@zauder.org

Đilda Pečarić, Miroslav Tuđman
Faculty of Humanities and Social Sciences
Department of Information Sciences
The University of Zagreb
Ivana Lučića 3, 10000 Zagreb, Croatia
dpecaric@ffzg.hr; mtudman@ffzg.hr

Abstract. *The paper discusses usage of citation indices and similar databases as the source of data for determining scientific professional advancement. A short study on per-author basis, using Web of Science, Scopus, Google Scholar, Croatian scientific bibliography and online catalogue of the National and University Library in Zagreb, has been undertaken and shows disparate results for usage of these resources in mentioned context. Various problems surrounding such studies are discussed. Data complementation and especially firmer, more interconnected, national databases are recommended.*

Authors primary recommendation is thus to connect current national sources in a more wholesome system providing quality access to bibliometric data in this context. Fragments of the system are already present and usable: Crosbi (scientific bibliography), NSK (national catalogue), Hrčak (journal full text database)). This would impact both the self-consciousness of the individual scientist and the scientific politics bypassing the colonial and provincial mentality imposed by the monopoly of "international" commercial databases. Such integrated model of wholesome representation and distribution of scientific production on national level would serve as an answer to problems inherent to usage of citation databases such as WoS and Scopus.

Keywords. science evaluation, citation indices, bibliometrics

1 Introduction

Scientists and scientific communities are under constant pressure. When science was considered to be the basic factor in social and economic growth they were threatened by the maxim „publish or perish". New, more subtle, methods of evaluation of scientific production were developed with ever increasing usage of Science citation index and other bibliographic, especially citation, databases.

In this respect, basic input for measuring the efficiency of a scientist is his "productivity" and "impact". Productivity usually operationalized through the number of items (in selected journals), and impact by citation counts in a certain body of literature as represented by citation indices. Journals are in turn evaluated by derived measures, most commonly the Thompson IF. Other derivatives, such as *h*-index, and analyses are sometimes used.

Presence in proprietary international databases (most notably Thompson Reuter's citation indices) is thus important not only for obtaining funding for scientific projects but also for professional advancement. According to administrative decision on national level, journals are divided in several groups. Legally, advancement is tied to A1 and A2 journals[1]. Papers published in international journals (A1) are automatically seen as providing scientific qualification, and citations in selective citation indices international status and respect. A2

[1] According to administrative decision A1 papers are papers that have one of these characteristics: a) journal or publication in which paper is published has international editorial board and international reviewers; b) journal or publication is in one data base conferred by National Science Council.
A2 papers are: a) paper published in journal and categorized as original scientific paper or review paper; b) chapter in the book; c) paper in the conference proceeding, if it is published entire article. (Pravilnik o uvjetima za izbor u znanstvena zvanja. Narodne novine, 2005).

papers overstep their provinciality only if they have "international review". It is a fact that the procedure of election for scientific and academic advancement (in Croatia) is brought to counting A1 and A2 papers. However, even this strict quantitative approach may not be brought under objective norms and standards.

Leaving aside, for the purposes of this article, the problems of our knowledge of what citation index data represent and what exactly are we measuring (Rip 1997), one of the major concerns lies in significant differences between arts, humanities and social sciences on one side and natural sciences, medicine and technology on the other (Nederhof 2006), as well as to more subtle differences between their subdisciplines. Arts, humanities and social sciences in conjunction with the nature of today's citation indices are especially problematic. The indices are primarily focused on journal items and partly on conference papers. English language is the main language of indexed publications, with other languages sparsely included but usually requiring bibliographic information in English.

Social sciences and humanities publications are thus often impaired from the start: books, chapters, local conferences and similar are not present in the citation data causing a lack of the possibility to operationalize authors' "productivity" and "impact" in an unbiased manner. These authors are fated for provincialisation.

In addition, the fact that the indices function on a commercial basis reduces the transparency of journal selection process as well as the nature or accuracy of algorithms preparing data or calculating various metrics.

The inability to apply standards and norms for evaluating scientific production in an unbiased manner causes more-or-less open conflict and frustration in scientific communities. The core of the problem is, in our opinion, that quantitative bibliometric indicators should not be the only measure of scientific quality. The goal of scientific thought is understanding and truth, which may not be reduced to normalization of scientific text production. The evaluation of scientific production through publication and citation counts and derived metrics is more in the function of knowledge industrialisation (to use K. P. Liessmann (2008) term) than science itself.

A frequent modern usage of citation indices is for per-author evaluation. This is also the use for which we believe these sources to be most problematic. We would like to illustrate this with a few examples.

2 Methodology

As the starting point, we chose the top 15 Croatian authors by citation from a previous research of 134 doctoral theses (having a total of 22.210 bibliographic items) defended in Croatian universities in the broader field of information sciences 1978-2007. (Đ. Pečarić, 2010.) This ensured authors of at least national significance and having various ties to information sciences.

The data about the authors' publications was gathered from international citation indices: Web of science (including SCI, SSCI, A&HCI), Scopus and Google Scholar. WoS and Scopus representing journal-level selective citation indices based on publisher metadata, and Scholar representing data obtained through web-indexing based, "anything goes" approach. Two sources of national importance were also included for comparison purposes: Croatian scientific bibliography[2] (CROSBI), and online catalogue of the National and University Library in Zagreb[3] (NSK). The former to provide insight into scholarly productivity from a national source, and the latter to show the possible broader publication scope of the included authors.

The searches yielded a total of 3753 bibliographic records from 4 of the sources, as Crosbi provides precise author identification and pre-aggregated per-author counts which were sufficient for this study.

It should be noted that the data is not provided with the intention to give insight into the productivity and citation of the authors themselves but rather to provide insight into the various relevant data sources and the more general problem of using bibliographic and citation databases for per-author analysis. Additionally, the search, preparation and analysis processes supply insights in the search process, data export/extraction and the quality of metadata obtained from various sources.

Being similar in nature, WoS and Scopus methodology was the same, excepting for some nuances in exact query structure as dictated by the sources. Items for each author were retrieved separately by performing simple author searches (allowing for variations of some authors names to ensure sufficient recall) and exported as tab

[2] http://bib.irb.hr
[3] http://katalog.nsk.hr

delimited text. Later analysis showed no overlap among the per-author sets.

While search strategy for Google Scholar was mostly the same, it required a more versatile handling of data as per-author searches yielded a total of 2246 records and Scholar does not allow any aggregated export or analytics. In this case, every item was downloaded in BibTeX format separately and subsequently parsed and aggregated per author. Unfortunately, Scholar does not include citation info in its exports so this was additionally done by hand.

Crosbi data was parsed from HTML (as it has no export possibilities), while online catalogue NSK allows MARC record exports from which the metadata used in the study was extracted. While parsing MARC to obtain items for bibliometric studies makes the process a lot harder, simpler exports were found not to contain enough data for quality study and might easily misinform. This is mainly due to the mapping between the standard notion of "author" and librarian notion of "responsibility", making it easy to confuse editor or translator with author.

3 Results

As Table 1. shows, most of the authors are represented with little to no items in selective citation indices. Being previously aware of problems of non-english speaking social science authors and WoS/Scopus data (Nederhof 2006), this was not surprising. Data obtained by WoS or Scopus complements each other and is also well complemented by other sources (Meho & Yang 2007). Some even found WoS to be irrelevant to their fields (Mingers & Lipitakis 2010).

Per-author sets yielded totals of 180 and 323 records for WoS and Scopus, respectively. After cleaning the sets by hand to include only relevant items, these numbers were reduced to 26 and 69, the main bulk of the difference between the indices being in few Croatian journals included in Scopus, not in WoS (Informatologia, Naše more). Relatively low precision of the search is mostly due to name synonyms (exacerbated by the fact that the names are ASCII only "surname, n." form, e.g. Tomislav Šola -> Sola, T), but also due to some search engine behaviour (e.g. novosel -> novosel'stev; plenkovic, m* -> plenkovic-moraj, a), and some common mistakes in author names (e.g. first surname interpreted as middle name).

Table 1. Productivity and citation in WoS/Scopus

author	corpus of phd theses	WoS		Scopus	
	total citations	n journal items (n citations)	n found items (n citations)	n journal items (n citations)	n found items (n citations)
Bauer, A.*	22	n/a	2313 (n/a)	n/a	3519 (n/a)
Lasić-Lazić, J.	29	0 (0)	0 (0)	1 (0)	1 (0)
Maroević, I.	36	0 (0)	0 (0)	3 (0)	3 (0)
Novosel, P.	53	2 (0)	2 (0)	0 (0)	13 (0)
Plenković, M.	68	0 (0)	0 (0)	12 (2)	12 (2)
Prelog, N.	21	0 (0)	0 (0)	0 (0)	0 (0)
Šola, T.	22	5 (1)	23 (50)	3 (2)	48 (819)
Srića, V.	55	1 (0)	14 (78)	1 (0)	19 (39)
Topolovec, V.	30	4 (29**)	4 (29)	6 (18**)	6 (18)
Tuđman, M.	54	3 (1)	3 (1)	3 (0)	3 (0)
Verona, E.	30	2 (1)	122 (1107)	0 (0)	171 (1163)
Vreg, F.	23	0 (0)	0 (0)	4 (0)	5 (0)
Zelenika, R.	24	6 (4)	6 (4)	30 (7)	30 (7)
Žiljak, V.	35	2 (2)	3 (3)	4 (5)	4 (5)
Žugaj, M.	25	0 (0)	3 (1)	2 (0)	7 (1)

* search yielding to many items to be usable
** citations are for papers in biology journals

These IR nuances may give quite erroneous information "at a glance". In this context, subsequent filtering of literature presupposes at least the knowledge of the author's field and no duplicate names in the same field or full text review. However, as some of the problems stem from the publisher metadata itself, some cases may simply miss information for this task. Presumptions about the exact field of the author may also be false as the authors may have more or less multidisciplinary interests during their careers. This leaves exact author identification a significant problem for this sources. As the table shows, author "Bauer, A" has a common name making search result set too broad for the data to be usable, especially using Scholar (see Table 2.).

Both databases have exporting capabilities in formats ready for analysis. Metadata is of high quality and information rich. Besides problems inherent in searching by author names and per-discipline significance, this makes data from these sources relatively easy to work with (in the technical sense).

Most of the so far mentioned is in contrast with Google Scholar data. Having a general, web-indexing approach, Scholar gathers a very diverse set of items. As Table 2. shows, the item and citation counts obtained from Google Scholar are far higher and show a part of publishing activity missed by WoS and Scopus.

Most of the overlap between the above data and WoS/Scopus data is in "journal items" and few "conference items" subsequently published in journals. Google data also shows many of the authors have "book" as highly productive and cited category illustrating the need for complementing WoS/Scopus data. Unfortunately, while Scholar data shows a richer publication scope, this makes it neither usable nor comparable with other sources.

Scholar data seems more like "random cuts from bibliographic metadata" than anything else. While it did find a very diverse set of items per author, its metadata was also of worst quality among the included sources. Mistakes included: often erroneous item type (defaulting to "article"; "chapter" almost nonexistent), mismatched field values, inclusion of translator or editor as the author, many duplicates, items with almost no data, items that refer to tables of contents having titles such as "original scientific paper", and so on.

Although the data has been cleaned by hand (often referencing external sources), due to the limited information it provides, type of item information should be interpreted in a more general manner than the other sources (for example, unlike for Crosbi and NSK data, "book" category includes textbooks). The lack of information per item also makes algorithmic approaches to handling this data difficult if not impossible. This contrasts with information rich items gained from WoS and Scopus as well as the data from the national catalogue.

Table 2. Google Scholar item and citation counts

autor	n books (n citations)	n book chapters (n citations)	n journal items (n citations)	n conference items (n citations)
Bauer, A.*	n/a	n/a	n/a	n/a
Lasić-Lazić, J.	6 (10)	6 (1)	42 (11)	42 (17)
Maroević, I.	20 (61)	17 (15)	90 (31)	33 (2)
Novosel, P.	8 (7)	7 (47)	14 (10)	1 (0)
Plenković, M.	14 (46)	29 (42)	142 (35)	72 (21)
Prelog, N.	1 (0)	2 (0)	11 (4)	2(0)
Šola, T.	6 (56)	8 (3)	45 (41)	24 (5)
Srića, V.	25 (130)	3 (4)	17 (23)	4 (4)
Topolovec, V.	0 (0)	2 (2)	15 (15)	4 (9)
Tuđman, M.	13 (46)	11 (4)	31 (16)	11 (4)
Verona, E.	14 (63)	1(0)	5 (56)	1 (6)
Vreg, F.	21 (129)	3 (12)	19 (23)	4 (2)
Zelenika, R.	67 (132)	5 (3)	122 (47)	27 (8)
Žiljak, V.	2 (2)	11 (8)	36 (16)	56 (40)
Žugaj, M.	14 (46)	8 (4)	38 (10)	28 (12)

* search yielding to many items to be usable (>12.000)

Table 3. Crosbi and online catalogue NSK

autor	CROSBI*					NSK	
	author books	book chapers	items in CC journals	items in other journals	proceedings items	author books	journal items
Bauer, A.	0	0	0	0	0	5	1
Lasić-Lazić, J.	4	6	0	10	40	3	11
Maroević, I.	9	11	0	69	5	15	65
Novosel, P.	0	2	0	4	3	12	5
Plenković, M.	5	31	2	77	43	7	35
Prelog, N.	0	1	0	2	0	4	22
Šola, T.	3	5	0	12	0	5	11
Srića, V.	2	2	0	0	0	29	26
Topolovec, V.	0	1	1	2	5	1	12
Tuđman, M.	2	13	1	6	7	8	10
Verona, E.	1	0	3	5	0	8	0
Vreg, F.	0	0	0	0	0	6	7
Zelenika, R.	7	4	0	93	35	9	148
Žiljak, V.	4	9	2	15	44	7	21
Žugaj, M.	5	4	0	17	20	6	51

* CROSBI categorization was inherited but some categories were combined to more general ones

Scholar proved highly sensitive to variations in author names. Skipping any variations in search proved to lose recall significantly, sometimes missing authors' most cited items. Other researchers noted many other inconsistencies in Scholar search process (Jacsó 2010).

Processing Google data took about 10 times more than all the other sources together. As this is in agreement with other Scholar studies (Meho & Yang 2007), we believe this source to be highly problematic for larger studies in this context.

National sources bypass the problem of per-author set identification, as they codify authors by unique IDs rather than name. They are, however, strictly bibliographic in nature, offering no citation data.

Crosbi data shows scientific text output per scientist, but with one possible source of inconsistency: Crosbi data is filled in by scientists themselves, frequently on a voluntary basis. NSK data shows the possibly broader publication scope of the authors and is, in this respect, most comparable with Google data. It should also be noted that NSK data covers a longer period, especially for books.

4 Discussion

4.1 Dataset comparison

Described data offers different insights into authors' production and impact from different angles and shows disparate results. While comparing or mapping these datasets may be the obvious choice to alleviate many of the problems, there are problems to this approach. Two main problems we identified in this respect are item count criteria and incomparable categories.

WoS is highly selective so one might simply count all the included items. But what to count in Scholar data as items vary from prestigious journals to lecture scripts? How many items are an item, the item's translation and the second edition? Different, possibly equally valid, choices will cause different counts per researcher and thus impact both direct usage of this data and usage for complementing the data from different sources. While every researcher may operationalize such research in own context, the larger problem of standardizing the process for unbiased per-author evaluation still remains.

Additionally, item categories from different sources make direct comparison of counts between different datasets difficult and possibly misleading. For example, "book" from Scholar is a different concept than Crosbi's several categories for "book" and the way MARC codifies publication types.

4.2 Citation indices usage and abusage

The usage of citation databases such as WoS, Scopus or Google Scholar as a resource indicating scientific productivity (and as the criterion for professional advancement) is based on the presumption that those databases enable empirical insight into knowledge maps and/or scientific intellectual structures. This presumption is based on many bibliometric studies of development of science and scientific disciplines, scholarly cooperation, scientific productivity, centres of scientific excelence, and similar, done during the last 30 years. Unfortunately, there are few studies showing the limits of usage of quantitative bibliometric indicators for evaluating and grading the knowledge value of scientific publications.

Even the data we have shown, based on a small number of authors, indicate the limitations of citation indices. On one hand, social sciences, humanities, natural sciences, medicine and technology are treated in different ways based on what "science" represents in different cultures and societies. Despite those differences, the journal is postulated as the dominating communication channel, which may not be true for all the sub-disciplines of social sciences and especially humanities. Besides domination of a single communication channel, one world language (English) is taken to be the primary language of science. By the nature of primary objects of their observance, this puts many disciplines (linguistics, literature, history, history of art, ethnology, communicology, etc.) on the margins of "global" citation databases. In the centre of the interest of these databases, and one should not forget they are primarily commercial in nature, is that part of scientific production which in one way or another has a technological or commercial use.

A large number of questions follows from just stated. We would like to warn about two unknowns. WoS and Scopus are two commercial databases most frequently used as a proof of citation of "local" authors, and thus of their participation and impact in "global" scientific communication. However, it is unknown to what degree are these databases used and how relevant are they as an information source for individual (especially humanities and social science) disciplines. It would be a paradox to use those databases for evaluating productivity (in humanities and social sciences) which are not used as an information source for scholarly research.

The other is a question of theoretical limits of bibliometric indicators and analyses offered by citation indices: does citation and co-citation analysis offer indicators of communicational or cognitive networks? If the scope of these analyses is primarily in describing scientific communication, these indicators are secondary for evaluation of cognitive networks (Đ. Pečarić, M. Tuđman, 2011.).

5 Conclusion

Citation indices are valuable for various bibliometric/scientometric purposes such as journal analysis, field/community analyses (as operationalised through per-journal selective process), insight into publishing or thematic trends, and for similar research.

However, they are problematic as resources for per-author evaluation. Included sources show quite varied data in this respect. WoS and Scopus rely mostly on journals, the former historically built around citation based metrics but also claiming for regional content: "Citation analysis may be applied but the real importance of the Regional journal is measured by the specificity of its content rather than its citation impact."[4] Unfortunately, it does not explain the proposed measure. Scholar offers a completely different set of advantages and problems as discussed, while national sources have either uncertain coverage and/or are not built for these purposes. They also do not provide any citation data.

While using these sources in a complementary manner is definitely a recommendation, it is not without its barriers and pitfalls. Besides the problems already described, using these sources in a complementary manner may simply be too time consuming for large scale usage, as the metadata is of diverse quality, codified in different ways and present in quite different formats.

Our primary recommendation is thus to connect current national sources in a more wholesome system providing quality access to bibliometric data in this context. Fragments of the system are already present and usable. For example, Crosbi (scientific bibliography), NSK (national catalogue), Hrčak (journal full text database)) but each provides only its own layer of information without communication between

[4] http://thomsonreuters.com/products_services/science/free/essays/regional_content_expansion_wos/

the layers and without providing an interface satisfying bibliometric needs.

A project which would methodologically connect already relevant sources should not have their institutional merging as a goal, but rather the valid representation on national level and promotion of scientific production on the international. Catalogue NSK represents the national production of all communication forms. CROSBI represents the scholarly production financed by the public funds, but does not include a critical verification of the data. Equal criteria for attribution of scholarly production could be insured with independent verification. This would impact both the self-consciousness of the individual scientist and the scientific politics bypassing the colonial and provincial mentality imposed by the monopoly of "international" commercial databases. Hrčak (journal full text database), would ensure the availability of scientific papers to the global public and thus full integration in global science. It could also serve as the basis for a national citation index.

Such integrated model of wholesome representation and distribution of scientific production on national level would serve as an answer to problems inherent to usage of citation databases such as WoS and Scopus. It would also reduce the frustrations growing from usage of "productivity", "impact", *h*-index, IF and other bibliometric indicators for evaluation of scientific efficiency. It would also provide the means for unbiased (or, at least, less biased) evaluation of national and global representation of scientific production.

References

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

[2] Liessmann K P: **Teorija neobrazovanosti. Zablude društva znanja.** Naklada Jesenski i Turk, Zagreb, 2008.

[3] Meho L I, Yang K: I**mpact of data sources on citation counts and rankings of LIS faculty: Web of science versus scopus and google scholar.** Journal of the American Society for Information Science and Technology, 58(13), 2007, pp. 2105-2125.

[4] Mingers J, Lipitakis E A E C G: **Counting the citations: a comparison of Web of Science and Google Scholar in the field of business and management.** Scientometrics, 85(2), 2010, pp. 613-625.

[5] Nederhof A J: **Bibliometric monitoring of research performance in the Social Sciences and the Humanities: A Review**. Scientometrics, 66(1), 2006, pp. 81-100.

[6] Pečarić Đ: **Razvoj informacijskih znanosti u Hrvatskoj. Bibliometrijska analiza doktorskih disertacija iz informacijskih znanosti 1978.-2007.** (Doctoral dissertation). Filozofski fakultet Sveučilišta u Zagrebu, Zagreb, 2010.

[7] Pečarić Đ, Tuđman M: **About the differences between communication networks and cognitive networks. Contribution to Research of Bibliometric methods in Information Science.** QQML2011: 3rd International Conference on Qualitative and Quantitative Methods in Libraries. Athens, Greece (manuscript), 2011.

[8] **Pravilnik o uvjetima za izbor u znanstvena zvanja**. Narodne novine, 84, 2005.

[9] Rip A: **Qualitative conditions of scientometrics: The new challenges**. Scientometrics, 38(1), 1997, pp.7-26.