The Web of Knowledge and Google Scholar: non-Western countries occurring in the byline versus country top level domain queries

The Stimulate 8 Group

Vrije Universiteit Brussel (VUB), Brussels, Belgium

The STIMULATE 8 Group consists of: Anne Sylvia ACHOM (Uganda), Helen Hagos BERHE (Ethiopia), Sangeeta Namdev DHAMDHERE (India), Alicia ESGUERRA (The Philippines), Nguyen Thi Ngoc HOAN (Vietnam), John KIYAGA (Uganda), Sheldon Miti MAPONGA (Zimbabwe), Yohannis MARTÍ-LAHERA (Cuba), Kelefa Tende MWANTIMWA (Tanzania), Marlon G. OMPOC (The Philippines), A.I.M. Jakaria RAHMAN (Bangladesh), Ronald ROUSSEAU (Belgium), Bahiru Shifaw YIMER (Ethiopia).

Abstract

This investigation illustrates differences between data available in professional databases such as the Web of Knowledge and data that are freely available on the Internet via Google Scholar. Our findings seem to indicate that, in general non-Western countries are better represented in the Web of Science than in Google Scholar. Through our results we illustrate one aspect of the digital divide between Western countries and other ones, in particular developing countries.

Keywords: digital divide; Web of Knowledge; Google Scholar; h-index; ideal world of science; topic searches

1. Introduction

STIMULATE stands for *Scientific and Technological Information Management in Universities and Libraries: an Active Training Environment*. It is an international training programme in information management, supported by the Flemish Interuniversity Council (VLIR), aiming at young scientists and professionals from developing countries. The programme has a dual purpose: it intends to develop the personal professional skills of the participants, and the participants are actively encouraged to transfer their newly acquired knowledge and skills to their colleagues and other stakeholders in their home country (Nieuwenhuysen & Vanouplines, 1997; Nieuwenhuysen, 2003; Stimulate 6 Group, 2007).

One of the higher level STIMULATE courses introduces students to the use of the World Wide Web and to bibliographic databases such as Thomson/Reuters's

Web of Knowledge as tools for library management and research evaluation (Stimulate 6 Group, 2007). This article is the result of the 'active training part' of this particular course. It illustrates differences between data available in professional databases such as the Web of Science and data that are freely available on the Internet via Google Scholar. In general non-Western countries are better represented in the Web of Science than in Google Scholar.

2. An ideal world of science

In an ideal world scientists publish their results in peer reviewed journals or conference proceedings and put a preprint version in a local, e.g. university, repository. The whole world of science is interconnected via the Internet.

An ideal publication-citation database covers all peer-reviewed journals and conference proceedings in the world.

An ideal search engine covers the whole Internet and clusters results such that each cluster refers to one item. More precisely: a cluster brings duplicate items in mirror sites together; preprint versions in an institutional repository and the published version on a publisher's website; etc..

In this ideal world the number of existing scientific publications, e.g. per country, as found in the ideal database is equal to the number of publications, e.g. per country, as found by the ideal search engine.

3. Aim

It would be a fine research project to study all differences between reality and the ideal world of science, and how this difference has changed over the years. Our investigation does not go that far, but has two rather modest aims: first, to draw attention to the existing gap between reality and the ideal (although some aspects, such as the underrepresentation of Third World countries, are well known), and second to investigate one aspect of this gap. Practically, for some topics, we compare the ranking of the top ten countries in the Web of Science (in short: WoS) with the ranking of these same ten countries based on the number of retrieved items (on this same topic) in Google Scholar. It is clear that in an ideal world these rankings must be the same. This study must be considered a pilot study for a larger and more thorough investigation.

4. Method

Each member of the STIMULATE 8 team chose a topic, preferably related to the country or the region he or she originated from. This topic was then represented by a word or phrase and with this word or phrase a topic search was performed in the Web of Science (December 2008). The five available databases were used:

Science Citation Index Expanded (data available from 1955 on), Social Science Citation Index (data since 1956), Arts & Humanities Citation Index (data since 1975), Conference Proceedings Citation Index – Science (data since 1990) and Conference Proceedings Citation Index – Social Science & Humanities (data since 1990).

The number of retrieved documents was noted and, via the 'citation report' option, also the h-index was obtained. This provided some background information which we will briefly discuss further on.

The essential part of our investigation consisted of using the "Analyze" option in order to obtain a ranking of countries that published (read: whose addresses occurred in the address line) about this topic. The ranked list was saved, and after some cleaning the top 10 countries (or more in case of ties) were kept. Cleaning means that we added the results for Germany, Federal Republic of Germany and the Democrat Republic of Germany, leading to a result for the two German countries (denoted as 'Germany', for short); and we obtained a representative number for the United Kingdom, based on the results for England, Scotland, Wales and North Ireland.

Next a search in Google Scholar aiming at retrieving the same topic (using the same query, or a slightly adapted one) was performed for each of the 10 countries. For example: the query Kiliman* in WoS became

(Kilimanjaro OR Kilimandjaro) AND site:tz

when we wanted scientific articles available on a Tanzanian website dealing with the topic Kilimanjaro. This search was repeated a total of ten times, once for each top 10 country (according to the WoS ranking). Results and details are shown in the next section and in the Appendix. No cleaning for duplicates was performed.

5. Data and some results

The following queries were performed in the WoS (Table 1). In case of differences the Google Scholar query is shown between square brackets. The total number of retrieved documents (in the WoS) and the topic h-index (Banks, 2006; the STIMULATE 6 Group, 2007) are also shown.

Some comments: although the AND-operator is unnecessary in Google and Google Scholar we used it anyway, as it made queries logically clear. The h-index for malaria was obtained by hand as the number of retrieved documents was too high for automatic determination. The terms (pesticide OR molluscicidal) were added to the query 'endod' as it turned out that Endod is also proper name. The term elephant* retrieved besides articles related to the well-known large

herbivorous animal with a long trunk, also quite some articles dealing with the elephant seal. However, we saw no reason to eliminate them. We kept the British term “diarrhoeal” as in this way a country such as Bangladesh entered the top 10. Results for the query on the Rwandan genocide and on endod are not used for further analysis as the first one included only two non-Western countries (so-called region B countries, see Section 6) and the second one did not retrieve enough results in Google Scholar.

Table 1

Query	# documents retrieved	h-index
(Rwanda* OR Ruanda*) AND genocide [(Rwanda OR Rwandan OR Ruanda OR Ruandan) AND genocide]	333	10
Pollution AND India	1448	35
Zambezi	411	26
(Vietnam OR “viet nam”) AND bay* [(Vietnam OR “viet nam”) AND (bay OR bays)]	104	12
Kiliman* [Kilimanjaro OR Kilimandjaro]	365	23
Pinatubo	1,359	77
Policosanol	213	32
Coffee AND arabica	1,257	35
Diarrhoeal	1,468	61
Ebola	1,340	76
Malaria	38,824	201
Elephant* [elephant OR elephants]	8,133	85
Stevia OR steviol	535	29
(endod AND (pestic* OR mollusc*)) OR (“phytolacca dodecandra”) [(endod AND (pesticide OR molluscicidal)) OR (phytolacca dodecandra)]	111	18

The complete results for each query, including details for each top 10 country, are given in the Appendix.

6. Findings

For simplicity we divided the world into two regions. Part A includes the Western industrialized countries, namely the USA, Canada, Australia, New Zealand and all European Countries (except Russia). All other countries are grouped together in region B.

Our main finding is that in the majority of cases (9 out of 12; two queries are not considered by lack of relevant data) B countries occupy a position which is farther away from the top position in Google Scholar than in the Web of Science. If Japan and South Africa had been placed in group A then the difference would have been even larger.

We note that our findings are based on a small sample and cannot be considered to give statistical evidence. Yet, based on this case study, there is no reason to think that the Internet offers a fairer representation of non-Western countries' science than the Web of Science.

We are fully aware of the fact that 'site'- searches in Google Scholar are not the same as searches for address fields in a bibliographic database. Yet, by concentrating on country sites we eliminate websites from publishers (all .com sites) and large international organizations (.org or .net sites). In this way, a country's scientific output, as it is made public by the country itself, is compared to another country's scientific output.

7. Some observations related to the h-index

As an aside we note the dependence of the h-index on the number of publications. A non-linear least-square fitting yields: $h = 2.44\ N^{0.416}$ ($R^2 = 0.91$), where N denotes the number of publications. This finding confirms earlier observations, see e.g. (Molinari & Molinari, 2008; STIMULATE 6 Group, 2007), in line with the Egghe-Rousseau model (Egghe & Rousseau, 2006) for the h-index.

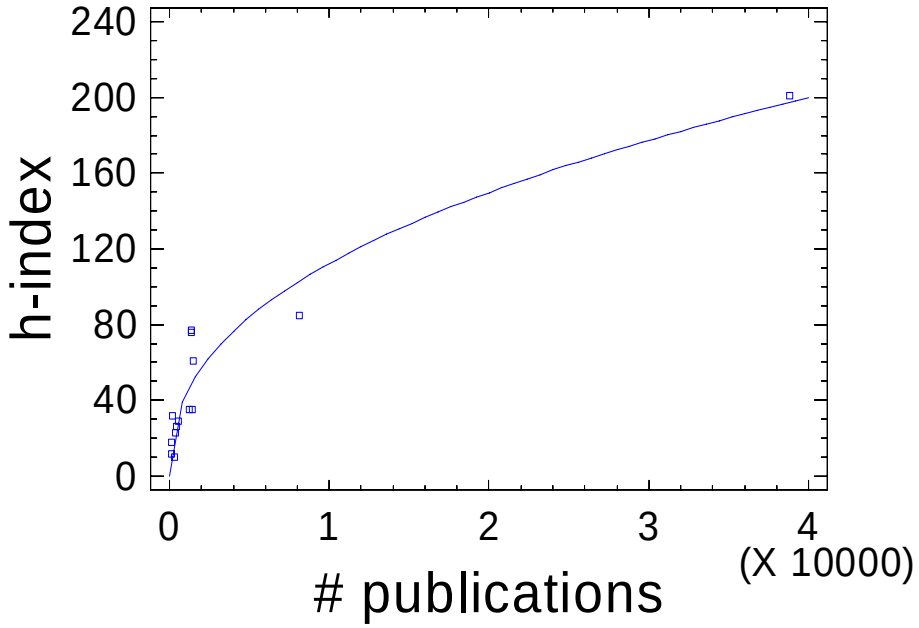

Figure 1. Relation between the h-index and the number of publications based on the data shown in Table 1.

8. Conclusion

It is sometimes thought that the Web of Science offers a Western view on science (Gibbs, 1995; Garfield, 1997; Kieling & Gonçalves, 2007), while the Internet is considered to be an equal playground, or at least a huge opportunity for equal treatment (Chan & Kirsop, 2001). Certainly, the Web of Science is largely based on Western journals and conference proceedings, yet a moment's reflection reveals that, for reasons of technological infrastructure the Internet is certainly not an equal playground (Oyelaran-Oyeyinka & Lal, 2005) and search engines show large biases against languages other than English (Aguillo et al., 2006). Note, however, that language bias does not play a role in our investigation as all queries are performed in English and target scientific articles included in the WoS.

Web presence inequality is confirmed by our investigation: with few exceptions (South Africa and Japan) the Web of Science as reflected by addresses, favours B countries' science more than the Internet, as covered by Google Scholar and measured using country codes (this is a huge caveat!). As such our article studies one aspect of the digital divide (Yu, 2006).

Leaving aside possible bias in the way Google Scholar views the Web (Vaughan & Thelwall, 2004), it seems that B countries, and especially the developing countries among this group, should do more to publicize scientific achievements of their scientists. Following others, e.g. (Chan & Kirsop, 2001), it is this group's opinion that all forms of Open Access can be a big step forward in this direction.

Finally, we hope that colleagues will take up the challenge and expand our preliminary findings. Based on our suggestions it should not be difficult to draw a plan for a research project mapping and explaining observed inequalities between Western countries (our region A), other industrialized countries (such as Japan, South Africa and China), and developing countries, in particular the Least Developed ones (LDCs).

Acknowledgement

The members of the STIMULATE 8 Group express their sincere thanks to Professor Paul Nieuwenhuysen (VUB, Brussels) and the VLIR (Flemish Interuniversity Council) who made this multinational collaboration possible. They thank Raf Guns (Univ. Antwerp) and Wolfgang Glänzel (K.U.Leuven) for useful suggestions improving the article.

References

I. F. Aguillo, B. Granadino, J.L. Ortega and J.A. Prieto (2006). Scientific research activity and communication measured with cybermetrics indicators. *Journal of the American Society for Information Science and Technology*, 57, 1296-1302.

M.G. Banks (2006). An extension of the Hirsch index: indexing scientific topics and compounds. *Scientometrics*, 69, 161-168.

L. Chan and B. Kirsop (2001). Open Archiving opportunities for developing countries: towards equitable distribution of global knowledge. *Ariadne*, 30. http://www.ariadne.ac.uk/issue30/oai-chan/intro.html

L. Egghe and R. Rousseau (2006). An informetric model for the Hirsch-index. *Scientometrics*, 69, 121-129.

E. Garfield (1997). A statistically valid definition of bias is needed to determine whether the Science Citation Index discriminates against third word journals. *Current Science*, 73(8), 639-641.

W.W. Gibbs (1995). Lost science in the Third World. *Scientific American*, 273(2), 76-83.

J. E. Hirsch (2005). An index to quantify an individual's scientific output. *Proceedings of the National Academy of Sciences of the United States of America*, volume 102, number 46, pp. 16569-16572.

C. Kieling and R. R. F. Gonçalves (2007). Assessing the quality of a scientific journal: the case of *Revista Brasileira de Psiquiatria. Revista Brasileira de Psiquiatria*, 29, 177-181.

J-F. Molinari and A. Molinari (2008). A new methodology for ranking scientific institutions. *Scientometrics*, 75, 163-174.

P. Nieuwenhuysen (2003). International training programs in Brussels related to scientific information and ICT. In: Proceedings of the Second Open Round Table on Developing countries access to scientific knowledge: quantifying the digital divide. Editors: Hilda A. Cerdeira and Enrique Canessa, Trieste : Abdus Salam International Center for Theoretical Physics, October 2003, 135 pp., ISBN 92-95003-22-5, pp. 35-38.

P. Nieuwenhuysen and P. Vanouplines (1997). International training courses on retrieval and management of information in science and technology. *Information Development*, 13(1), 23-26.

B. Oyelaran-Oyeyinka and K. Lal (2005). Internet diffusion in sub-Saharan Africa: a cross-country analysis. *Telecommunications Policy*, 29, 507-527.

STIMULATE 6 Group (2007). The Hirsch index applied to topics of interest to developing countries. *First Monday,* volume 12, number 2, available at http://firstmonday.org/issues/issue12_2/stimulate/index.html

L. Vaughan and M. Thelwall (2004). Search engine coverage bias: evidence and possible causes. *Information Processing and Management,* 40, 693-707.

LZ. Yu (2006). Understanding information inequality: making sense of the literature of the information and digital divides. *Journal of Librarianship and Information Science*, 38, 229-252.

Appendix: Detailed results for each query.

The tables shown below contain a short description of the query (not the full version) in the first column, the top-10 countries (sometimes more in case of ties) in terms of published articles on this topic, according to the WoS (in the second column); the corresponding number of retrieved documents (third column); the number of documents retrieved in Google Scholar on this same topic, using the country top level domain as part of the query (column 5), and the ranking of the these countries according to the number of documents retrieved in Google Scholar (column 4).

We also give the Pearson correlation coefficient between columns 3 and 5, and the Pearson rank correlation coefficient between columns 2 and 4. Finally average rank of A countries and B countries for each of the rankings is given.

Query: pollution AND India	Wos ranking	WoS number of documents	GS ranking	GS number of sites
India	1	1079	2	2270
USA	2	192	1	6900
United Kingdom	3	56	3	1830
Japan	4	55	7	1040
Germany	5	42	5	1290
China	6	37	9	643
Netherlands	7	22	8	910
Canada	8	20	4	1640
Bangladesh	9	17	11	27
France	10.5	15	6	1180
Philippines	10.5	15	10	101
Spearman rank correlation	0.72			
Pearson correlation	0.27			
Average rank A countries	5.92		4.50	
Average rank B countries	6.10		7.80	

Query: Zambezi	Wos ranking	WoS number of documents	GS ranking	GS number of sites
Zimbabwe	1	92	8	16
USA	2	82	1	584
United Kingdom	3	65	3	154
South Africa	4	55	2	257
Zambia	5	41	10	8
Germany	6	27	4	125
Australia	7	24	6	70
Botswana	8	15	9	12
Netherlands	9	14	5	100
France	10	13	7	64
Spearman rank correlation	0.33			
Pearson correlation	0.48			
Average rank A countries	6.17		4.33	
Average rank B countries	4.50		7.25	

Query: Vietnam AND bay*	Wos ranking	WoS number of documents	GS ranking	GS number of sites
Vietnam	1	42	12	56
USA	2	22	1	3290
Russia	3	17	14	42
Japan	4	14	3	287
Belgium	5	8	10	90
France	7.5	5	6	116
Italy	7.5	5	7	113
Thailand	7.5	5	13	50
UK	7.5	5	5	271
Canada	12	4	2	600
Denmark	12	4	9	89
Germany	12	4	4	285
Philippines	12	4	11	67
Sweden	12	4	8	97
Spearman rank correlation	-0.13			
Pearson correlation	0.27			
Average rank A countries	8.61		5.78	
Average rank B countries	5.50		10.60	

Query: Kiliman*	Wos ranking	WoS number of documents	GS ranking	GS number of sites
Tanzania	1	99	7	56
USA	2	86	1	679
United Kingdom	3	52	3	164
Norway	4	47	5	107
Germany	5	31	2	184
Kenya	6	27	9	6
Austria	7.5	11	8	31
France	7.5	11	4	128
South Africa	9	8	6	80
Nigeria	10	7	10	0
Spearman rank correlation	0.53			
Pearson correlation	0.53			
Average rank A countries	4.86		3.57	
Average rank B countries	6.50		8.00	

Query: Pinatubo	Wos ranking	WoS number of documents	GS ranking	GS number of sites
USA	1	686	1	883
Germany	2	164	3	246
United Kingdom	3	137	5	132
Japan	4	116	2	266
France	5	105	4	203
Canada	6	76	6	115
Italy	7	70	10	45
Russia	8	51	9	53
Philippines	9	47	7	89
Switzerland	10	37	8	80
Spearman rank correlation	0.83			
Pearson correlation	0.98			
Average rank A countries	4.86		5.29	
Average rank B countries	7.00		6.00	

Query: policosanol	Wos ranking	WoS number of documents	GS ranking	GS number of sites
Cuba	1	132	2	62
USA	2	39	5.5	4
Italy	3	8	7.5	3
Germany	4	6	5.5	4
Canada	5	5	3	10
China	6	4	1	75
Argentina	8.5	3	9	1
Japan	8.5	3	7.5	3
Netherlands	8.5	3	10	0
South Korea	8.5	3	4	5
Spearman rank correlation	0.44			
Pearson correlation	0.53			
Average rank A countries	4.50		6.30	
Average rank B countries	6.50		4.70	

Query: coffee AND arabica	Wos ranking	WoS number of documents	GS ranking	GS number of sites
Brazil	1	304	2	1670
France	2	176	4	343
USA	3	143	3	1290
Costa Rica	4	85	1	3820
Germany	5	84	6	261
United Kingdom	6	70	5	311
Kenya	7	57	10	2
Mexico	7	54	8	94
Japan	9	50	7	153
India	10	48	9	63
Spearman rank correlation	0.81			
Pearson correlation	0.32			
Average rank A countries	4.00		4.50	
Average rank B countries	6.33		6.17	

Query: diarrhoeal	Wos ranking	WoS number of documents	GS ranking	GS number of sites
United Kingdom	1	307	2	359
USA	2	280	1	598
Bangladesh	3	160	9	11
India	4	137	5	168
Australia	5	80	3	345
Nigeria	6.5	51	10	1
Sweden	6.5	51	7	116
Germany	8.5	49	6	135
Switzerland	8.5	49	8	53
South Africa	10	48	4	251
Spearman rank correlation	0.40			
Pearson correlation	0.66			
Average rank A countries	4.60		4.20	
Average rank B countries	5.88		7.00	

Query: ebola	Wos ranking	WoS number of documents	GS ranking	GS number of sites
USA	1	716	1	896
Germany	2	161	2	307
France	3	119	3	248
United Kingdom	4	87	5	121
Canada	5	83	4	162
Japan	6	75	6	107
Russia	7	62	9	13
Belgium	7	53	8	37
Congo (+ Zaire)	9	47	10	0
Switzerland	10	36	7	91
Spearman rank correlation	0.90			
Pearson correlation	0.97			
Average rank A countries	5.25		5.00	
Average rank B countries	7.33		8.33	

Query: malaria	Wos ranking	WoS number of documents	GS ranking	GS number of sites
USA	1	11,310	1	8860
United Kingdom	2	7,429	4	2070
France	3	3,456	2	6090
Australia	4	2,099	5	1840
Germany	5	2,082	3	3080
Switzerland	6	1,990	8	811
Thailand	7	1,679	9	355
India	7	1.638	6.5	1180
Kenya	9	1,480	10	22
Netherlands	10	1,362	6.5	1180
Spearman rank correlation	0.80			
Pearson correlation	0.77			
Average rank A countries	4.43		4.21	
Average rank B countries	8.00		8.50	

Query: elephant*	Wos ranking	WoS number of documents	GS ranking	GS number of sites
USA	1	2911	1	11700
United Kingdom	2	816	2	2440
South Africa	3	450	9	833
Germany	4	438	6	1520
Australia	5	374	4	1950
Canada	6	373	3	2190
Brazil	7	339	7	1170
India	8	247	10	470
France	9	239	5	1590
Japan	10	210	8	1140
Spearman rank correlation	0.55			
Pearson correlation	0.98			
Average rank A countries	4.50		3.50	
Average rank B countries	7.00		8.50	

Query: stevia OR steviol	Wos ranking	WoS number of documents	GS ranking	GS number of sites
Japan	1	103	3	135
USA	2	92	4	49
Brazil	3	73	1	287
Mexico	4	47	2	137
Argentina	5.5	27	7	29
India	5.5	27	8	12
Thailand	7	24	9	5
Canada	7	21	6	31
UK	9.5	19	10	2
Germany	9.5	18	5	34
Spearman rank correlation	0.68			
Pearson correlation	0.60			
Average rank A countries	7.25		6.25	
Average rank B countries	4.33		5.00	

The following two queries did not yield sufficient data and were eliminated from the analysis. For the first one only two B countries figured in the top 10, for the second one not enough results were obtained in Google Scholar.

Query: Rwandan genocide	Wos ranking	WoS number of documents	GS ranking	GS number of sites
USA	1	122	1	2120
United Kingdom	2	51	4	477
Belgium	3	15	8	96
Norway	4	10	9	90
Canada	5	9	3	493
France	6.5	8	7	131
Rwanda	6.5	8	10	7
Netherlands	8	6	6	209
South Africa	9.5	4	2	494
Australia	9.5	4	5	324
Spearman rank correlation	0.08			
Pearson correlation	0.92			
Average rank A countries	4.88		5.38	
Average rank B countries	8.00		6.00	

Query: endod	Wos ranking	WoS number of documents	GS ranking	GS number of sites
USA	1	32	1	4
Ethiopia	2	21	8.5	0
Denmark	3	15	4	1
Switzerland	4	13	4	1
Zimbabwe	5	10	8.5	0
Norway	6	6	4	1
Egypt	7	4	8.5	0
South Africa	7	4	4	1
Swaziland	9	4	8.5	0
Germany	10.5	3	4	1
Kenya	10.5	3	8.5	0
Spearman rank correlation	0.34			
Pearson correlation	0.69			
Average rank A countries	4.90		3.40	
Average rank B countries	6.92		7.75	

Remark

We notice sometimes large differences between the Spearman rank correlation coefficient and the Pearson correlation coefficient. Such differences are not exceptional: the Spearman correlation is less susceptible to outliers or other values that may influence the results. Remember also that these two correlation coefficients are used to test different hypotheses. Pearson tests a linear relation, while Spearman tests a monotonic relation.