Short communication

Strange attractors in the Web of Science database

Miguel A. García-Pérez *

Departamento de Metodología, Facultad de Psicología, Universidad Complutense, Campus de Somosaguas, 28223 Madrid, Spain

ARTICLE INFO	ABSTRACT
	Accurate computation of h indices or other indicators of research impact requires access to databases supplying complete and accurate citation information. The Web of Science (WoS) database is widely used for this purpose and it is generally deemed error-free. This note describes an inaccuracy that seems to affect differentially non-English sources and targets in WoS, namely, "phantom citations" (i.e., papers reported by WoS to cite some article when they actually did not) and their concentration around particular articles that are thus dubbed "strange attractors". The analysis of references in (and citations to) papers in two English sources and two non-English sources reveals that phantom citations and other errors of indexing occur about twice as often with non-English items. These and other errors of commission affect about 1% of the cited references in the WoS database, and they may reveal large-scale problems in the reference matching algorithm in WoS. © 2010 Elsevier Ltd. All rights reserved.

1. Introduction

Sources providing complete and accurate citation records are necessary for an assessment of the impact of research at the individual, institutional, or national level. Thomson-Reuter's Journal Impact Factor and Hirsch's (2005) *h* index are alternative indicators of research performance which are both computed from citation counts. Thomson Reuter's *Web of Science* (WoS) is the citation-enhanced database most widely used for the accrual of citation records that are, in turn, used to compute these indicators, but other platforms are also available which generally provide different citation records for the same target papers (Bar-Ilan, 2008; Jacso, 2008a, 2008b, 2008c, 2008d; Levine-Clark & Gil, 2009; Meho & Rogers, 2008; Meho & Yang, 2007; Norris & Oppenheim, 2007; Vaughan & Shaw, 2008; Vieira & Gomes, 2009; Walters, 2009). Differences in the length and content of the lists of citation records provided by different platforms are understandable as a result of differences in database coverage, and this is one of the threats to an accurate and complete accrual of citation records (Frandsen & Nicolaisen, 2008; Glavel & Iselid, 2008).

Indeed, citation-enhanced databases only retrieve citations from papers included in their own databases, which potentially results in a substantial number of what Jacso (2008a, 2008b) termed "orphan" references, that is, references that do not count as citations because the database does not have a master record for the cited item. But a complete accrual of citation records is also jeopardized by what Jacso (2008a, 2008b) termed "stray" references, which are references that do not count as citations either because an error of referencing by the citing authors precludes linking the cited paper to a master record that actually exists in the database. Further threats to the accurate retrieval of citation records come from errors of commission by data-entry operators (i.e., when accurate references in articles are introduced with error in the database, yielding also stray references of a different origin; or when accurate references in articles are not introduced at all in the database) and from the presence of "phantom citations" (i.e., papers tagged in the database as citing an article that they

* Tel.: +34 913 943 061; fax: +34 913 943 189.
  E-mail address: miguel@psi.ucm.es.

actually did not cite; Jacso, 2008a). Errors of commission are deemed infrequent in for-fee databases. Yet, a recent study involving a sample of 157 journal articles for which WoS listed 1335 citations revealed that about 0.3% of those citations were phantom (García-Pérez, in press).

Despite their low prevalence, phantom citations in WoS seem to repeatedly affect specific articles that we dub "strange attractors." This paper reports this peculiarity, which differentially affects non-English sources and authors. Our goal is not to present a thorough analysis of the WoS database so as to provide accurate figures of the prevalence of the problem but, rather, to describe and document it so that users are aware of it and, hopefully, WoS staff manages to fix whatever causes the problem.

2. Strange attractors

For an example of the type of error that will be described in this paper, Fig. 1 shows the articles that WoS reports as citing the target paper by GarcíaFerrando indicated at the top. Yet, none of these articles actually cited GarcíaFerrando's paper, as is evident upon inspection of the reference lists in the purported citing articles. A comparison of the actual reference lists with those in the WoS record for the purported citing papers reveals that GarcíaFerrando's paper has always taken the place of a reference to the same book chapter (namely, García-Pérez, M.A., Eye movements and perceptual multistability. In E. Chekaluk & K.R. Llewellyn (Eds.) *The Role of Eye Movements in Perceptual Processes*. Amsterdam: Elsevier, 1992, 73–109). In other words, WoS assigns five phantom citations to GarcíaFerrando's paper.

Papers reported by WoS as cited in such a large number of articles that have not actually cited them are referred to as "strange attractors" here, because it is unlikely that such accumulation of phantom citations has occurred accidentally. For a further example, Fig. 2 shows another list of articles reported by WoS as citing the target paper by García-Pérez at the top. The three papers indicated by arrows represent phantom citations too, making the target paper another strange attractor. In this case, the actual items that the impostor replaces in the WoS list of cited references varies across the purported citing papers (see Table 1).

These two cases involve non-English targets or sources. To investigate this characteristic, the cited reference lists in the WoS records of all papers published in 2008 in the non-English source *Revista de la Facultad de Agronomía de la Universidad*

Fig. 1. List of citing articles in WoS. None of the five purported citing papers actually cited GarcíaFerrando's paper.

Fig. 2. List of citing articles in WoS. None of the three purported citing papers indicated with arrows actually cited García-Pérez's paper.

del Zulia (RFAUZ) was checked against the actual reference lists in those papers, and a similar analysis was carried out for all papers published in 2008 in the English source *Spatial Vision* (SV). The results are shown in Table 2.

Although both journals published about the same number of papers in 2008, SV papers included almost twice as many references as RFAUZ papers (1417 vs. 764). WoS records included more than 99% of the references in either case, but the percentage of linked references (i.e., those matching a master record in the database) was substantially larger for SV than for

Table 1

Actual articles cited in the papers indicated by arrows in Fig. 2 (#1, #3, and #7). These articles are found in the reference lists of the actual papers but they are missing in their WoS records. For unknown reasons, their place in those records was taken by the impostor paper by García-Pérez which none of the papers actually cited.

#1	García-Zamora, P., Ros, R. M. & Guerra, J. (2000). Vegetación briofítica de las sierras de Filabres, Cabrera, Alhamilla y Cabo de Gata (Almería, SE de España). *Cryptogamie Bryologie*, *21*, 19–75.
#3	García Peris P. & Velasco Gimeno C. (2007). Evolución en el conocimiento de la fibra. *Nutrición Hospitalaria*, *22*(Suppl. 2), 20–25.
#7	García Winder, M. (2007). Factores que afectan la competitividad del sector avícola latinoamericano. InterCambio. Políticas, Comercio y Agronegocios. Director de Área de Desarrollo de los Agronegocios. Instituto Interamericano de Cooperación para la Agricultura. http://infoagro.net/comercio/intercambio/avicola.pdf.

Table 2

Comparison of Revista de la Facultad de Agronomía de la Universidad del Zulia (RFAUZ) and Spatial Vision (SV) in 2008.

	RFAUZ	SV
Number of published papers	35	33
Number of references across all papers	764	1417
Number of references included in the WoS records (%)	757 (99.08)	1408 (99.36)
Number of linked references in the WoS records (%)	200 (26.18)	834 (58.86)
Number of missing references in the WoS records (%)	7 (0.92)	9 (0.64)
Number of phantom citations in the WoS records (%)	4 (0.52)	0 (0.00)
Number of errors of commission across cited reference lists in the WoS records (%)	3 (0.39)	11 (0.78)

Table 3

Actual cited papers and impostors involved in the three remaining phantom citations across RFAUZ papers.

Actual cited paper	Impostor (paper that took the place of the actual cited paper in the WoS record)
Martínez Sáez, S., Hernández, J. & Guevara, R. (2000). Determinación relativa de fenoles, flavonoides y esteroides en cinco leguminosas tropicales. *Revista de Producción Animal, 12*, 37–39.	Martín, F. & Kearney, J. F. (2000). Positive selection from newly formed to marginal zone B cells depends on the rate of clonal production, CD19, and *btk*. *Immunity, 12*(1), 39–49.
Shuler, K. D. (1999). Bell pepper variety trial results. Southwest Florida Research & Education Center. University of Florida. http://www.imok.ufl.edu/LIV/groups/cultural/trial/pepper.htm.	Shulman, L. S. (1999). Taking learning seriously. *Change, 31*(4), 11–17.
Ventura, G. & Jimenez, R. (2004). Evaluación de sistemas de siembra y distancia entre plantas en la producción orgánica de banano (*Musa* AAA cv. Gran enano) en República Dominicana. pp. 23–31. In: J. Orozco, M. Orozco, R. Zapata, A Vizcaína, A. Morfin & J. Hernández (Eds.). *Memorias XVI Reunión Internacional de ACORBAT*, Oaxaca, México.	Ventura, I. (2004). *Quaestiones* and encyclopedias: Some aspects of the late medieval reception of the pseudo-Aristotelian *problemata* in encyclopedic and scientific culture. pp. 23–42. In: A. A. MacDonald & M. W. Twomey (Eds.). *Schooling and Society: The Ordering and Reordering of Knowledge in the Western Middle Ages*. Groningen Studies in Cultural Change, Vol. 6. Leuven, Belgium: Peeters.

RFAUZ (58.86% vs. 26.18%). Seven references (0.92%) actually included in RFAUZ papers were missing in their WoS records, compared to 9 (0.64%) in SV papers. However, the WoS records of RFAUZ papers included four phantom citations, compared to none in SV. One of these was shown in Fig. 2; the remaining three and the items that they replaced are listed in Table 3.

The cause of phantom citations and their concentration around strange attractors is hard to trace, but they seem to affect non-English sources or targets preferentially. The impostors seem to be papers whose bibliographical details bear some resemblance to the items they replace: identical publication year and volume or initial page number combined with some similarity in first author's name. However, an explanation based on similarity is untenable because actual errors of referencing by the authors generally render stray references even when these errors are minimal (see Table S1 in the Supplemental Information for sample cases). Stray references also occur when accurate bibliographical data in the citing paper are erroneously introduced in WoS records (a typical outcome for non-English sources; see Osca-Lluch, Civera Mollá, & Peñaranda Ortega, 2009), or when master records contain incorrect bibliographical data. Three such errors of commission were identified in the cited reference lists of the WoS records of RFAUZ papers, and 11 for SV papers (see Table S2 in the Supplemental Information). All of these errors along with the missing references indicated in Table 2 deprive papers from legitimate citations that WoS should report. Other errors of commission were also found in our analysis which were not tallied because they involved items in sources not covered by WoS.

Errors of commission and phantom citations are the exclusive responsibility of the service provider. When they are both added up, the picture that emerges is that 1.83% (14/764) of the total number of references across RFAUZ papers published in 2008 were introduced with error in WoS, whereas the corresponding figure for SV is 0.78% (11/1417), or about 60% less.

Strange attractors and phantom citations are not rare encounters in WoS according to the preceding figures and the author's own experience. To further illustrate, citations to papers published in *Psicothema* (a Spanish source) and in the *Journal of Mathematical Psychology* (JMP; an English source) during the years 2001 and 2002 were retrieved from WoS. A comparison of the reference lists in the purported citing papers with the WoS records yielded the results shown in Table 4. WoS listed

Table 4

Comparison of *Psicothema* and *Journal of Mathematical Psychology* (JMP) in WoS as of May 7, 2010.

	Psicothema		JMP	
	2001	2002	2001	2002
Number of papers[a]	94	154	40	35
Number of papers with non-zero citation counts	64	98	37	29
Largest citation number	14	17	72	113
Second largest citation number	13	15	30	21
Total number of citations	216	380	306	333
Number of phantom citations across both years (%)	8 (1.34)		2 (0.31)	

[a] The count of JMP papers excludes 14 book reviews in 2001 and 15 in 2002.

639 citations to 75 JMP papers, and two of those citations (0.31%) were phantom; in contrast, WoS listed 596 citations to 248 *Psicothema* papers, and eight of those citations (1.34%) were phantom. Table S3 in the Supplemental Information lists these phantom citations, revealing also three additional strange attractors. In other words, English sources seem less affected by errors of commission in WoS records.

3. Conclusion

Phantom citations in WoS are not rare, particularly for non-English sources and authors. For unknown reasons, some papers in the database become strange attractors and accumulate phantom citations (Figs. 1 and 2), depriving other papers and authors from their legitimate citations. These misdeeds merge with other errors of commission that render stray references in the WoS database in a way that affects all indicators of research performance or impact that are based on citation counts, including *h* indices. It might be argued that this problem should not affect *h* indices, given that Rousseau (2007) developed a theoretical argument whereby the *h* index is robust to missing citations; yet, an empirical study (García-Pérez, in press) has shown that the *h* index is not that robust in real conditions. The exact magnitude and consequences of phantom citations and strange attractors in WoS is hard to ascertain, but the misdemeanor encourages the use of other platforms for the accrual of complete citation records (see also García-Pérez, in press).

Acknowledgments

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.joi.2010.07.006.

References

Bar-Ilan, J. (2008). Which *h*-index?—A comparison of WoS, Scopus and Google Scholar. *Scientometrics*, *74*(2), 257–271.

Frandsen, T. F., & Nicolaisen, J. (2008). Intradisciplinary differences in database coverage and the consequences for bibliometric research. *Journal of the American Society for Information Science and Technology*, *59*(10), 1570–1581.

García-Pérez, M. A. (in press). Accuracy and completeness of publication and citation records in the Web of Science, PsycINFO, and Google Scholar: A case study for the computation of *h* indices in Psychology. *Journal of the American Society for Information Science and Technology*, doi:10.1002/asi.21372.

Glavel, Y., & Iselid, L. (2008). Web of Science and Scopus: A journal title overlap study. *Online Information Review*, *32*(1), 8–21.

Hirsch, J. E. (2005). An index to quantify an individual's scientific research output. *Proceedings of the National Academy of Sciences of the USA*, *102*(46), 16569–16572.

Jacso, P. (2008a). Testing the calculation of a realistic *h*-index in Google Scholar, Scopus, and Web of Science for F. W. Lancaster. *Library Trends*, *56*(4), 784–815.

Jacso, P. (2008b). The pros and cons of computing the *h*-index using Google Scholar. *Online Information Review*, *32*(3), 437–452.

Jacso, P. (2008c). The pros and cons of computing the *h*-index using Scopus. *Online Information Review*, *32*(4), 524–535.

Jacso, P. (2008d). The pros and cons of computing the *h*-index using Web of Science. *Online Information Review*, *32*(5), 673–688.

Levine-Clark, M., & Gil, E. L. (2009). A comparative citation analysis of Web of Science, Scopus, and Google Scholar. *Journal of Business and Finance Librarianship*, *14*(1), 32–46.

Meho, L. I, & Rogers, Y. (2008). Citation counting, citation ranking, and *h*-index of human-computer interaction researchers: A comparison of Scopus and Web of Science. *Journal of the American Society for Information Science and Technology*, *59*(11), 1711–1726.

Meho, L. I., & Yang, K. (2007). Impact 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), 2105–2125.

Norris, M., & Oppenheim, C. (2007). Comparing alternatives to the *Web of Science* for coverage of the social sciences' literature. *Journal of Informetrics*, *1*(2), 161–169.

Osca-Lluch, J., Civera Mollá, C., & Peñaranda Ortega, M. (2009). Consecuencias de los errores en las referencias bibliográficas, El caso de la revista Psicothema. *Psicothema*, *21*(2), 300–303.

Rousseau, R. (2007). The influence of missing publications on the Hirsch index. *Journal of Informetrics*, *1*(1), 2–7.

Vaughan, L., & Shaw, D. (2008). A new look at evidence of scholarly citation in citation indexes and from web sources. *Scientometrics*, *74*(2), 317–330.

Vieira, E. S., & Gomes, J. A. N. F. (2009). A comparison of Scopus and Web of Science for a typical university. *Scientometrics*, *81*(2), 587–600.

Walters, W. H. (2009). Google Scholar search performance: Comparative recall and precision. *Portal: Libraries and the Academia*, *9*(1), 5–24.