Opinion Paper

The museum of errors/horrors in Scopus

Fiorenzo Franceschini*, Domenico Maisano, Luca Mastrogiacomo

Politecnico di Torino, DIGEP (Department of Management and Production Engineering), Corso Duca degli Abruzzi 24, 10129 Torino, Italy

ARTICLE INFO

ABSTRACT

Recent studies have shown that the Scopus bibliometric database is probably less accurate than one thinks. As a further evidence of this fact, this paper presents a structured collection of several weird typologies of database errors, which can therefore be classified as *horrors*. Some of them concern the incorrect indexing of so-called *Online-First* paper, duplicate publications, and the missing/incorrect indexing of references. A crucial point is that most of these errors could probably be avoided by adopting some basic data checking systems.

Although this paper does not provide a quantitative and systematic analysis (which will be provided in a future publication), it can be easily understood that these errors can have serious consequences such as: (i) making it difficult or even impossible to retrieve some documents, and (ii) distorting bibliometric indicators/metrics relating to journals, individual scientists or research institutions.

Our attention is focused on the Scopus database, although preliminary data show that the Web of Science database is far from being free from these errors. The tone of the paper is deliberately provocative, in order to emphasize the seriousness of these errors.

© 2015 Elsevier Ltd. All rights reserved.

1. Introduction

In the last decade, bibliometric databases have come into the life of scientists, being commonly used for: (i) searching scientific documents, (ii) providing information on the impact of the scientific output of individuals and/or research institutions, and (iii) supporting the selection of scientific journals in which to publish.

The abundance of bibliometric disciplinary databases (e.g., PubMed in Biology, Medicine and Life Sciences, Mathematics in MathSciNet, PsycINFO in Psychology, IEEEXplore in Engineering, EconLit in Economics, etc.) contrasts with the relatively limited number of multidisciplinary databases: Scopus, Web of Science (WoS), and perhaps Google Scholar (GS), although the latter is something between a search engine and a bibliometric database. A peculiarity of GS is to fully automatically index publications/citations through web crawlers; this inevitably causes many errors, sometimes very grotesque (Labbé, 2010), and (almost) completely disqualifies GS with respect to its two competitors (i.e., Scopus and WoS), to the extent that most consider GS simply as a search engine, certainly not a serious bibliometric database.

On the other hand, Scopus and WoS have a reputation for being relatively accurate, which puts them in a position of superiority with respect to GS, despite the latter indexes a significantly higher number of scientific documents (De Winter, Zadpoor, & Dodou, 2014). For this reason, (part of) the scientific community in some countries takes the information contained in Scopus and WoS as gospel (DORA, 2013; Harzing, 2013). Also, Scopus and WoS are accessible only under the payment of

* Corresponding author.
E-mail addresses: fiorenzo.franceschini@polito.it (F. Franceschini), domenico.maisano@polito.it (D. Maisano), luca.mastrogiacomo@polito.it (L. Mastrogiacomo).

rather expensive subscriptions, while GS is free. Although there is some information available on the (presumed) criteria adopted by Scopus and WoS for selecting/excluding scientific journals (Scopus Elsevier, 2015; Thomson Reuters, 2015), there is no public information about (i) the process for periodically including the latest articles into the database, (ii) the type of (meta)data received from publishers, (iii) the way these articles are linked with the relevant cited/citing paper, etc.

Being recognized as serious multidisciplinary databases, Scopus and WoS have a significant normative power. For example, their (not very transparent) policies about indexing or excluding journals and conference proceedings lead to inevitable discriminations between the documents indexed (i.e., the "good" ones) and those excluded (i.e., the "bad" ones); also, these databases regularly publish indicators – such as the *Impact Factor* (IF) by WoS or *SCImago Journal Rank* (SJR) and *Source Normalized Impact per Paper* (SNIP) by Scopus – that are commonly (mis)used for various bibliometric evaluations (Simons, 2008; DORA, 2013).

In the last couple of years, we have been investigating the Scopus and WoS errors, analyzing the so-called *omitted citations* – i.e., missing links between citing and cited papers in a database – which represent the major consequence of database errors (Franceschini, Maisano, & Mastrogiacomo, 2013). The most interesting result – which corroborates the findings of previous studies (Moed, 2002; Buchanan, 2006) – is that the omitted-citation rate of the two databases is not negligible: about 5% for Scopus and about 7% for WoS (Franceschini, Maisano, & Mastrogiacomo, 2015a); this means that at least 1/20 of the citations purportedly indexed by these databases are omitted! In addition, we showed that the editorial style of some publishers can favour database errors (Franceschini, Maisano, & Mastrogiacomo, 2014), and that, although the Scopus and WoS databases tend to be more and more careful in indexing new papers – probably taking advantage of the recent efforts by reviewers/editors/publishers in checking and correcting possible inaccuracies in the cited-article lists – they do little to correct the errors already present in the database (Franceschini, Maisano, & Mastrogiacomo, 2015b).

During our investigation we came across many weird errors, which are probably the result of fairly embarrassing negligence of databases. Despite being aware of the inevitable imperfection of all things in this world (including bibliometric databases!) and that *errare humanum est*, such negligence often made us doubt about the real level of care in data indexing/management.

In order for the reader to understand and share our concerns, we will introduce him/her into the *museum of horrors* (not just errors!) *in Scopus*, consisting of two rooms: (i) an *antechamber* containing some preliminary horrors, which are to some extent justifiable, and (ii) a *main hall* containing some unjustifiable horrors, sometimes exceeding the human imagination. The presentation of these horrors will follow a Rossini *crescendo*.

The decision to limit our analysis to the Scopus database comes from the fact that, on average, this database is supposed to be more accurate than WoS, and it is therefore not unreasonable to expect from it a certain rigour in terms of data accuracy/cleaning (Franceschini et al., 2015a,b).

Fig. 1. Example of missing citation link between a paper of interest (P_1) and a citing one (P_2), due to a typo concerning the title of P_1, in the cited-article list of P_2. The Scopus database was queried in August 2015.

2. Antechamber

The following two subsections describe two typologies of at least partly justifiable bibliometric-database horrors.

2.1. No error detection?

Inaccuracies in the cited-article list of a (citing) paper – e.g., concerning author name(s), issue year, title, journal name, volume, number, pagination, etc. – may sometimes compromise the correct determination of the link with cited papers. In the example in Fig. 1, this error is caused by a typo concerning the title of a paper of interest (P_1), which is reported in the list of another (citing) paper (P_2).

Paper of interest (P_1):

DOI: 10.1364/OE.16.007323

Cited-article list:

Abstract: The cylindrical coordinate machining method (CCM) is
[...]

©2008 Optical Society of America

OCIS codes: (220.1250) Aspherics; (220.1920) Diamond machining

References and links

1. X. Jiang, P. Scott, D. Whitehouse, "Freeform surface characterisation – A fresh strategy," Ann. CIRP **56**, 553-556 (2007).
2. H. N. Hansen, K. Carneiro, H. Haitjema, L. De Chiffre, "Dimensional micro and nano metrology," Ann. CIRP **55**, 721-743 (2006).
3. L. De Chiffre, H. Kunzmann, G. N. Peggs, D. A. Lucca, "Surfaces in precision engineering, microengineering and nanotechnology," Ann. CIRP **52**, 561-578 (2003).
4. Y. Takeuchia, S. Maedaa, T. Kawaib, K. Sawadab, "Manufacture of multiple-focus micro Fresnel lenses by means of nonrotational diamond grooving," Ann. CIRP **51**, 343-346 (2002).
5. C. C. Chen, C. M. Chen, J. R. Chen, "Toolpath generation of diamond shaping of aspheric lens array," J. Mate. Proc. Tech. **192-193**, 194-199 (2007).
6. H. B. Wu, Z. Q. Wang, R. L. Fu, J. Liu, "Design of a hybrid diffractive/refractive achromatized telecentric f-θ lens," Optik **117**, 271-276 (2006).
7. L. L. Doskolovich, N. L. Kazanskiy, S. I. Kharitonov, P. Perlo, S. Bernard, "Designing reflectors to generate a line-shaped directivity diagram," J. Mode. Opti. **52**, 1529-1536 (2005).
8. C. F. Cheung, L. B. Kong, W. B. Lee, "Modelling and simulation of freeform surface generation in ultra-precision raster milling," J. Engi. Manu. **220**, 1787-1801 (2006).
9. L. Piegl, W. Tiller, "The NURBS Book," (New York: Springer-Verlag, 1997).
10. W. Ma, J. P. Kruth, "NURBS curve and surface fitting for reverse engineering," J. Adva. Manu. Tech. **14**, 918-927 (2005).
11. Y. L. Ma, W. T. Hewitt, "Point inversion and projection for NURBS curve and surface: control polygon approach," Comp. Aide. Geom. Desi. **20**, 79-99 (2003).
12. K. H. Jeong, J. Kim, L. P. Lee, "Biological inspired artificial compound eyes," Science **312**, 557-561 (2006).

1. Introduction

Freeform surfaces can be used in optical systems to achieve novel functions, improve
[...]

References (not) reported by Scopus:

ISSN: 10944087   Source Type: Journal   Original language: English

DOI: 10.1364/OE.16.007323   PubMed ID: 18545437   Document Type: Article

Publisher: Optical Society of American (OSA)

References **??**

Fig. 2. Example of occult references, which are completely ignored by the Scopus database. Database was queried in August 2015.

In this specific case, Scopus was unable to identify and correct this inaccuracy. This could certainly be done by taking into account the other redundant data, available from the reference to P_1 (e.g., volume, number, pagination), which would make P_1 uniquely identifiable. Not surprisingly, WoS was able to correctly identify the link between P_1 and P_2.

This error typology is relatively frequent for both Scopus and WoS (García-Pérez, 2011; Olensky, Schmidt, & Eck, 2015).

2.2. Occult references

In this error typology, the cited-article list of the paper of interest is completely ignored by the database. For the purpose of example, let us consider the paper (P_1) with DOI: 10.1364/OE.16.007323, published by the journal "Optics Express" in 2008. The Scopus database does not report any of the (twelve) references of P_1, omitting the corresponding citations. This is probably due to the fact that the cited-article list of P_1 is positioned right below the abstract and not at the end of the document (see Fig. 2). Although we are aware that this editorial convention is quite unusual, it is curious that Scopus ignores the fact that a *research paper* – not just a *letter to the editor* or a *short communication* – has (apparently) no references.

3. Main hall

Having seen the serious but at least partly understandable errors/horrors contained in the antechamber, we now enter the main hall of the museum, which contains evidences of various unpardonable horrors.

3.1. No spell checker?

This error typology regards database transcription errors. In the example in Fig. 3, Scopus erroneously replaces the word "manufacturing", in the title of a paper of interest (P_1), with "manufacturingg". We believe that this specific error is quite serious, since it does not concern the spelling of a complicated acronym or author name, but just a common word of the English language. Probably, it could be avoided by the use of a simple spell-checker! It was found that this type of error can even compromise the attribution of the citations to P_1, obtained by citing papers reporting the title of P_1 correctly in their cited-article lists.

3.2. Citation infanticide

For several recent years now, scientific journals have been struggling to include the new-entry papers in their websites as soon as possible, in the form of so-called *Online-First* papers, i.e., papers not yet in the official version, but already available to the scientific community (Haustein, Bowman, & Costas, 2015). Apart from encouraging the spread of new knowledge, this mechanism allows journals to artificially extend the time-window for citation accumulation, resulting in a probable increase of the journal IF and other bibliometric indicators (Falagas & Alexiou, 2008).

Paper of interest (P_1):

DOI: 10.1243/09544054JEM1794

Title: Combining business process and failure modelling to increase yield in electronics manufacturing

Corresponding Scopus record:

Fig. 3. Example of database transcription error concerning the title of a paper (P_1). The Scopus database was queried in August 2015.

Paper of interest (P_1):
DOI: 10.1007/s10845-009-0341-3
Online-First availability date: 28 October 2009
Official Publication date: 2012

Citing paper (P_2):
DOI: 10.1016/j.ress.2011.09.008

Citation by P_2, obtained by the Online-First version of P_1:

[25] Wang X, Rabiei M, Hurtado J, Modarres M, Hoffman P. A probabilistic-based airframe integrity management model. Reliability Engineering & System Safety 2009;94:932–41.
[26] Guan X, Jha R, Liu Y. Probabilistic fatigue damage prognosis using maximum entropy approach. Journal of Intelligent Manufacturing 2009;1–9, doi:10.1007/s10845-009-0341-3.
[27] Tierney L, Kadane J. Accurate approximations for posterior moments and marginal densities. Journal of the American Statistical Association 1986;81:82–6.

Missing link by Scopus:

○ Wang, X., Rabiei, M., Hurtado, J., Modarres, M., Hoffman, P.
25 A probabilistic-based airframe integrity management model
(2009) *Reliability Engineering and System Safety*, 94 (5), pp. 932-941. Cited 30 times.
doi: 10.1016/j.ress.2008.10.010
POLITO SFX βετα [View at Publisher]

○ Guan, X., Jha, R., Liu, Y.
26
(2009) *Journal of Intelligent Manufacturing*, pp. 1-9. Cited 13 times.
10.1007/s10845-009-0341-3
POLITO SFX βετα ← missing link to P_1

○ Tierney, L., Kadane, J.
27
(1986) *Journal of the American Statistical Association*, 81, pp. 82-86. Cited 548 times.
POLITO SFX βετα [View at Publisher]

Fig. 4. Example of citation infanticide: a citation obtained by the Online-First version of paper P_1 (issued in October 2009) is lost after the publication of the relevant official version (in 2012). The Scopus database was queried in August 2015.

Bibliometric databases are also struggling to index Online-First papers as soon as possible. Unfortunately, the "double stage" of these papers can favour the generation of database errors; the most common is that of losing the citations obtained by the Online-First version of the paper of interest (P_1), after the publication of the relevant official version. In other words, the "infant" citations, obtained by the Online-First version of P_1, are "killed" when the official version is issued! See the example in Fig. 4.

These rather serious and frequent typology of errors make us raise a question (Haustein et al., 2015): is it reasonable that a database struggles to index the Online-First papers, if it does not have the ability to manage/update them properly?

3.3. Identity cloning

The DOI code is the ID of a scientific paper and allows to uniquely identify it. It is therefore paradoxical that multiple scientific papers share the same DOI code. Despite this, the Scopus database contains papers that (apparently) do it. The example in Fig. 5 shows that Scopus has "cloned" the DOI of the paper of interest (P_1), attributing it to another paper (P_2).

Paper of interest (P_1):

Title: Effects of gas composition during plasma modification of polyester fabrics
DOI: 10.1016/j.jmatprotec.2005.11.008

Paper with cloned DOI (P_2):

Title: Diversidade clínica e laboratorial no haplótipo bantu da anemia falciforme
Correct DOI: 10.1590/S1516-84842006000100010

Wrong DOI (according to Scopus): 10.1016/j.jmatprotec.2005.11.008
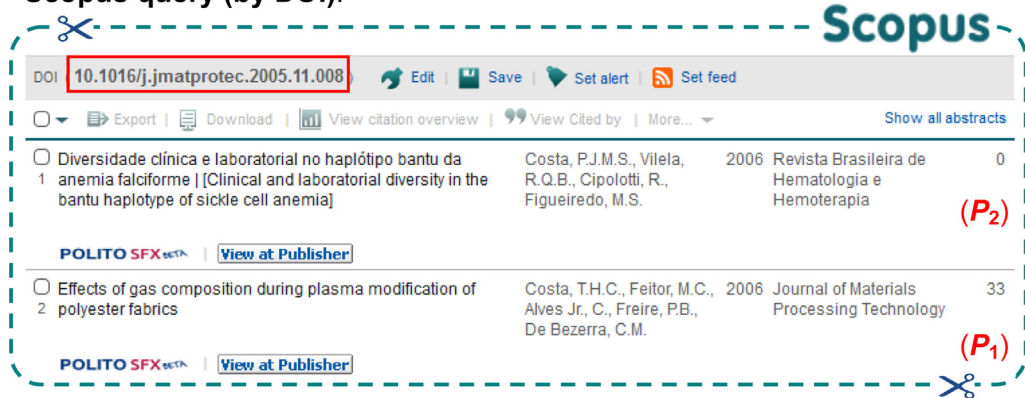
Scopus query (by DOI):

Fig. 5. Example of identity cloning of the paper P_1 into P_2. The Scopus database was queried in August 2015.

This error is even more horrible, considering that it could be prevented through a trivial check of the uniqueness of DOI codes.

3.4. Dissociative identity disorder

This error typology is represented by the simultaneous presence in the database of two distinct records concerning the same paper. This rather frequent duplication problem has recently been described by Valderrama-Zurián, Aguilar-Moya, Melero-Fuentes, and Aleixandre-Benavent (2015). This "dissociative disorder" can have different causes, such as: (1) missing removal of the *Online-First* version of a paper when indexing the official version (see the example in Fig. 6), (2) journal title variations (see the example in Fig. 7), (3) journal name changes, etc.; for details see Valderrama-Zurián et al. (2015).

Paper of interest (P_1):

Title: Influence of omitted citations on the bibliometric statistics of the major Manufacturing journals
DOI: 10.1007/s11192-015-1583-9

Scopus Query (by title):

Fig. 6. Example of simultaneous indexing of the same paper (P_1) in two distinct versions (i.e., the Online-First and the official one). The Scopus database was queried in August 2015.

Paper of interest (P_2):

Title: Importance of pollinators in changing landscapes for world crops
DOI: 10.1098/rspb.2006.3721

Scopus Query (by title):

Fig. 7. Example of simultaneous indexing of the same paper (P_2) in two distinct records, due to a journal title variation; adapted from (Valderrama-Zurián et al., 2015). It is surprising to notice that the record that attracted more citations (i.e., 721 against 141) is the one related to the alternative journal title, not the official one. The Scopus database was queried in October 2015.

Paper of interest (P_1):

Title: Estimation of noise covariance matrices for periodic systems
DOI: 10.1002/acs.1255

(True) cited-article list of P_1:

Cited-article list of P_1, according to Scopus:

Fig. 8. Example of citation tumour affecting the cited-article list of a paper of interest (P_1), according to Scopus. Database was queried in August 2015.

This type of error can produce significant distortions in the bibliometric indicators related to the authors and the institutions participating in the papers with duplicated records. Apart from contributing to fictitiously increase the number of papers, duplicate records may contribute to distort the number of citations that actually belong to a single paper. Precisely, Valderrama-Zurián et al. (2015) remark that both duplicates can have equal number of citations, different number of citations (see the example in Fig. 7), or one of the duplicates can be cited whilst the other not.

3.5. Citation tumours

In some cases, the list of a certain paper may contain more references than the correct ones, as if it was subject to an abnormal (tumoral) growth. Let us consider the example in Fig. 8, in which the paper of interest (P_1) contains a list of 24 papers; this list is eight-upled by the Scopus database, which reports 192 references, 168 (i.e., 192–24) of which are imaginary!

The most obvious consequence of this abominable error is the production of a large amount of so-called "phantom" citations—i.e., fictitious citations in favour of papers not actually cited by the paper of interest (Jacsó, 2006).

The type of tumour exemplified is *benign*, as the (168) phantom citations are added to the (24) true ones. We also came across the more dangerous *malignant* citation tumours, in which the phantom citations may even infiltrate the true ones, replacing (some of) them.

Paper of interest (P_1):

Title: Changeable Manufacturing - Classification, Design and Operation
DOI: 10.1016/j.cirp.2007.10.003

Partially "disintegrated" cited-article list of P_1 (by Scopus):

Fig. 9. Example of disintegration of several references (no less than 105, i.e. from reference number 45 to 149!).

3.6. Disintegration of references

In this error typology, the database transcribes the same number of references reported on the paper of interest (P_1); however, some of them are empty or – more picturesquely – have "disintegrated" (see the example in Fig. 9 in which 105 out of 152 total references have disintegrated). As a consequence, the citations given by P_1 to the disintegrated papers are lost.

4. Conclusions

This paper presented some typologies of weird errors/horrors of the Scopus database. Many of them, such as that concerning the duplication of DOI codes (*identity cloning*) or the insane modification of the cited-article list (*citation tumours* and *disintegration of references*), even exceed human imagination.

These horrors are even more surprising in view of the fact that: (i) Scopus is probably the most accurate multidisciplinary bibliometric database currently available (Franceschini et al., 2015a,b), and (ii) many of them could have been avoided right from the beginning, by adopting some basic data checking systems, e.g., identifying multiple documents with the same DOI code and/or title.

A limitation of this study is that it does not include neither a statistical analysis of these error/horror typologies, nor a systematic investigation of the relevant causes. Regarding the future, we plan to carry out a structured comparison of Scopus and WoS databases, based on a quantitative analysis of the error/horror typologies. Preliminary data show that these horrors are far from isolated cases, since they involve about 1% of the purportedly indexed citations, for both the Scopus and WoS database.

Although we are aware of the important role played by Scopus for the scientific community, we remark that these horrors can have serious consequences such as: (i) making it difficult or even impossible to retrieve some documents, and (ii) distorting bibliometric indicators/metrics relating to journals, individual scientists or research institutions.

Scopus (and WoS) should certainly put more effort into developing suitable tools for improving their data accuracy . . . at least as much as the effort they usually put in marketing campaigns!

References

Buchanan, R. A. (2006). Accuracy of cited references: The role of citation databases. *College & Research Libraries, 67*(4), 292–303.
De Winter, J. C., Zadpoor, A. A., & Dodou, D. (2014). The expansion of Google Scholar versus Web of Science: a longitudinal study. *Scientometrics, 98*(2), 1547–1565.
DORA. (2013). *San Francisco declaration on research assessment,*. ⟨http://am.ascb.org/dora/⟩ [20 August 2015].
Falagas, M. E., & Alexiou, V. G. (2008). The top-ten in journal impact factor manipulation. *Archivum immunologiae et therapiae experimentalis, 56*(4), 223–226.
Franceschini, F., Maisano, D., & Mastrogiacomo, L. (2013). A novel approach for estimating the omitted-citation rate of bibliometric databases. *Journal of the American Society for Information Science and Technology, 64*(10), 2149–2156.
Franceschini, F., Maisano, D., & Mastrogiacomo, L. (2014). Scientific journal publishers and omitted citations in bibliometric databases: Any relationship? *Journal of Informetrics, 8*(3), 751–765.
Franceschini, F., Maisano, D., & Mastrogiacomo, L. (2015a). Influence of omitted citations on the bibliometric statistics of the major manufacturing journals. *Scientometrics, 103*(3), 1083–1122.
Franceschini, F., Maisano, D., & Mastrogiacomo, L. (2015b). On the correction of "old" omitted citations by bibliometric databases. In *Proceedings of the 15th International Society of Scientometrics and Informetrics (ISSI) conference* , ISBN 978-975-518-381-7 (pp. 1200–1207).
García-Pérez, M. A. (2011). Strange attractors in the Web of Science database. *Journal of Informetrics, 5*(1), 214–218.
Harzing, A. W. (2013). Document categories in the ISI Web of Knowledge: Misunderstanding the social sciences? *Scientometrics, 94*(1), 23–34.
Haustein, S., Bowman, T. D., & Costas, R. (2015). When is an article actually published? An analysis of online availability, publication, and indexation dates. In *Proceedings of the 15th International Society of Scientometrics and Informetrics (ISSI) Conference* , ISBN 978-975-518-381-7 (pp. 1170–1179).
Jacsó, P. (2006). Deflated, inflated and phantom citation counts. *Online Information Review, 30*(3), 297–309.
Labbé, C. (2010). Ike Antkare, one of the great stars in the scientific firmament. *ISSI Newsletter, 6*(2), 48–52.
Moed, H. (2002). The impact-factors debate: The ISI's uses and limits. *Nature, 415*(6873), 731–732.
Olensky, M., Schmidt, M., & Eck, N. J. (2015). Evaluation of the citation matching algorithms of CWTS and iFQ in comparison to the Web of science. *Journal of the Association for Information Science and Technology*, http://dx.doi.org/10.1002/asi.23590
Scopus Elsevier. (2015). *Content policy and selection,*. ⟨https://www.elsevier.com/solutions/scopus/content/content-policy-and-selection⟩ [10 November 2015].
Simons, K. (2008). The misused impact factor. *Science, 322*(5899), 165.
Thomson Reuters. (2015). *The Thomson Reuters journal selection process,*. ⟨http://wokinfo.com/essays/journal-selection-process/⟩ [10 November 2015].
Valderrama-Zurián, J. C., Aguilar-Moya, R., Melero-Fuentes, D., & Aleixandre-Benavent, R. (2015). A systematic analysis of duplicate records in Scopus. *Journal of Informetrics, 9*(3), 570–576.