POLITECNICO DI TORINO

DIGEP
(Dept. of Management and Production Engineering)

2nd International Summer School on Scientometrics - "Modern Trends In Science: Scientometrics"

26-29 September 2018, Yerevan (Armenia)

# Data quality in bibliometric databases: remarks from different research perspectives

Domenico Maisano

# A similar contribution has been presented in:

Workshop: “Research evaluation in Italy”, 5th June 2018, Rome

- Giovanni Abramo, CNR (National Research Council), Italy;
- Jonathan Adams, Director of ISI, Clarivate Analytics, UK/USA;
- Dag W. Aksnes, NIFU, Norway.
- Cinzia Daraio, Università di Roma “La Sapienza”, Italy;
- Domenico A. Maisano, Politecnico di Torino, Italy.

https://clarivate.savoinspire.com/italia/

# Bibliometric database errors

| Error type | Pre-existing errors | Database mapping errors |
|---|---|---|
| Definition | Errors made by authors/editors/publishers when preparing the list of cited papers for their publication. | Failures to establish an electronic link between a cited paper and the corresponding citing papers that can be attributed to a data-entry error. |
| Examples | -Errors in the author name(s),<br>-Errors in paper title,<br>-Errors in issue year,<br>-Errors in volume number,<br>-Errors in pagination. | -Transcription errors,<br>-Target-source article record errors,<br>-Cited article omitted from a cited-paper list,<br>-Reason unknown. |
| Stage: | Production of the paper | Indexing |
| Parties involved: | Authors/editors/publishers | Database staff |

Buchanan, R.A. (2006). Accuracy of Cited References: The Role of Citation Databases. College & Research Libraries, 67(4): 292-303.

# Examples

## Example of *pre-existing error*

**Cited paper ($P_1$)**:

Authors: J. Dong, P.M. Ferreira, J.A. Stori
Title: Feed-rate optimization with jerk constraints for generating minimum-time trajectories
Source: International Journal of Machine Tools and Manufacture, 47(12-13): 1941-1955
DOI: 10.1016/j.ijmachtools.2007.03.006

**Citing paper ($P_2$)**:

Authors: X. Broquere, D. Sidobre, I. Herrera-Aguilar
Title: Soft motion trajectory planner for service manipulator robot
Source: IEEE/RSJ International Conference on Intelligent Robots and Systems, 2008. IROS 2008.
DOI: 10.1109/IROS.2008.4650608

**Reference to $P_1$ (with inaccurate author names), in the list of $P_2$**:

[8] P. F. Jingyan Dong and J. Stori, "Feed-rate optimization with jerk constraints for generating minimum-time trajectories," *International Journal of Machine Tools and Manufacture*, 2007.

**Reference to $P_1$ (with inaccurate author names), according to Scopus**:

Scopus

Kyriakopoulos, Konstantinos J., Saridis, George N.
7 MINIMUM JERK PATH GENERATION.
(1988) , pp. 364-369. Cited 72 times.
ISBN: 0818608528
POLITO SFX BETA View at Publisher

Jingyan Dong, P.F., Stori, J.
8 Feed-rate optimization with jerk constraints for generating minimum-time trajectories
(2007) *International Journal of Machine Tools and Manufacture*. Cited 2 times.

missing link to $P_1$

## Example of *database mapping error*

**(Citing) paper of interest ($P_1$)**:

Authors: J. Hong, D. Xu, P. Gong, J. Yu, H. Ma, S. Yao
Title: Covalent-bonded immobilization of enzyme on hydrophilic polymer covering magnetic nanogels
Source: Microporous and Mesoporous Materials, 109(1-3): 470-477
DOI: 10.1016/j.micromeso.2007.05.052

**Original list of ($P_1$)**:

**References**

[1] K.M. Koeller, C.H. Wong, Nature 409 (2001) 232.
[2] R. Sharma, Y. Chisti, U.C. Banerjee, Biotechnol. Adv. 19 (2001) 627.
[...]
[37] S. Rauf, A. Ihsan, K. Akhtar, M.A. Ghauri, M. Rahman, M.A. Anwar, A.M. Khalid, J. Biotechnol. 121 (2006) 351.
[38] S. Tembe, M. Karve, S. Inamdar, S. Haram, J. Melo, S.F. D'Souza, Anal. Biochem. 349 (2006) 72.

**Missing list in WoS**:

WEB OF SCIENCE™

Covalent-bonded immobilization of enzyme on hydrophilic polymer covering magnetic nanogels
By: Hong, J (Hong, J.); Xu, D (Xu, D.); Gong, P (Gong, P.); Yu, J (Yu, J.); Ma, H (Ma, H.); Yao, S (Yao, S.)
MICROPOROUS AND MESOPOROUS MATERIALS
Volume: 109 Issue: 1-3 Pages: 470-477
DOI: 10.1016/j.micromeso.2007.05.052
Published: MAR 1 2008

Citation Network
47 Times Cited
0 Cited References
View Citation Map
Create Citation Alert
(data from Web of Science ™ Core Collection)

absence of references

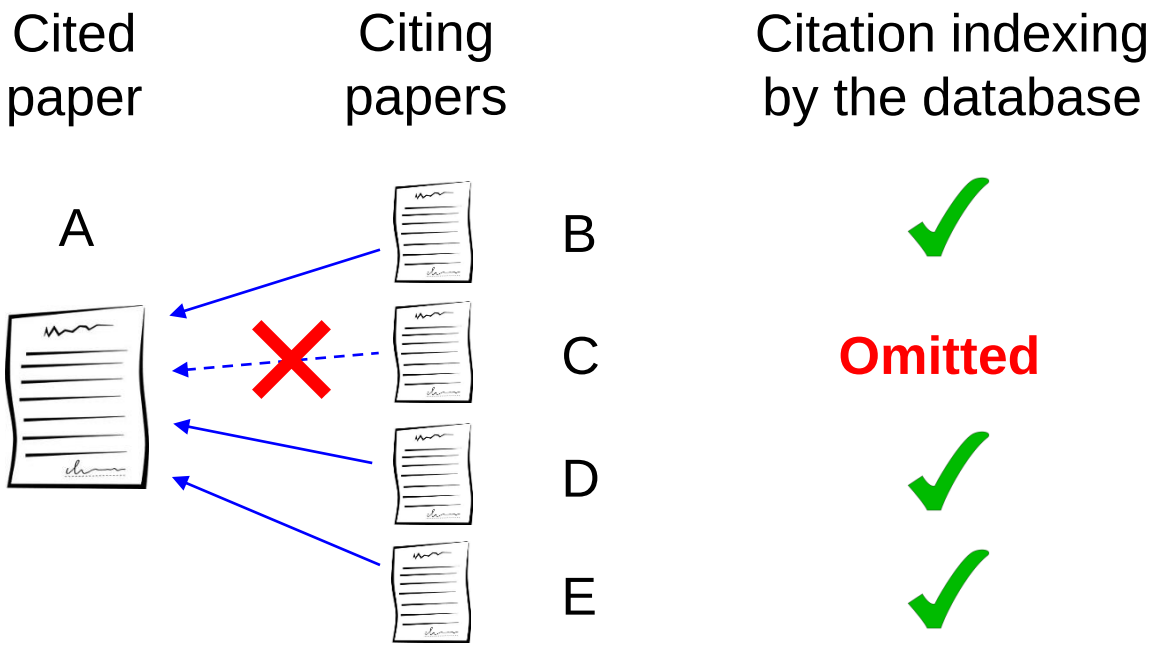
Franceschini, F., Maisano D., Mastrogiacomo L. (2016) Empirical analysis and classification of database errors in Scopus and Web of Science. Journal of Informetrics, 10(4): 933-953.

# Omitted citations (1)

- Bibliometric databases are not free from errors;
- Omitted citations represent one of the main consequences.

→ i.e., citations that should be ascribed to a certain (cited) paper but, for some reason, are lost.

# Omitted citations (2)




Citation statistics relating to paper A

| | |
|---|---|
| No. of **citations returned** by the database: | **3** |
| "True" no. of citations: | 4 |
| No. of citations **omitted** by the database: | **1** |
| **Omitted-citation rate**: | ***p* = 25%** |

# Automated algorithm (1)

- An automated algorithm for estimating the omitted-citation rate of databases has been developed.
- It requires the combined use of two (or more) bibliometric databases.
- *Hypothesis*: the mismatch between the citations occurring in one database and another one is evidence of possible errors/omissions.

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.

# Automated algorithm (2)

| Cite No. | Database 1 (DB1) | Database 2 (DB2) |
|---|---|---|
| 1 | ✓ | N/A (source **not** covered) |
| 2 | N/A (source **not** covered) | ✓ |
| **3** | ✗ | ✓ |
| **4** | ✓ | ✓ |
| **5** | ✓ | ✓ |
| **6** | ✗ | ✓ |
| 7 | N/A (source **not** covered) | ✓ |
| **8** | ✓ | ✗ |

Omitted-citation rate

(omitted cites/TO cites)

$p_{DB1} = 2/5 = 40\%$

$p_{DB2} = 1/5 = 20\%$

Theoretically Overlapping (TO) citations: citations coming from sources purportedly covered by both databases (5 citations in this case).

# Statistical model

- It allows to estimate the "true" value (with a suitable confidence interval) of the (bibliometric) indicators of interest, compensating for omitted citations.

$$C^* = \frac{C}{1-p} \pm 2 \cdot \sqrt{C \cdot p}$$

Total citations

$$IF^* = \frac{IF}{(1-p)} \pm 2 \cdot \sqrt{\frac{IF \cdot p}{P_{cit.}}}$$

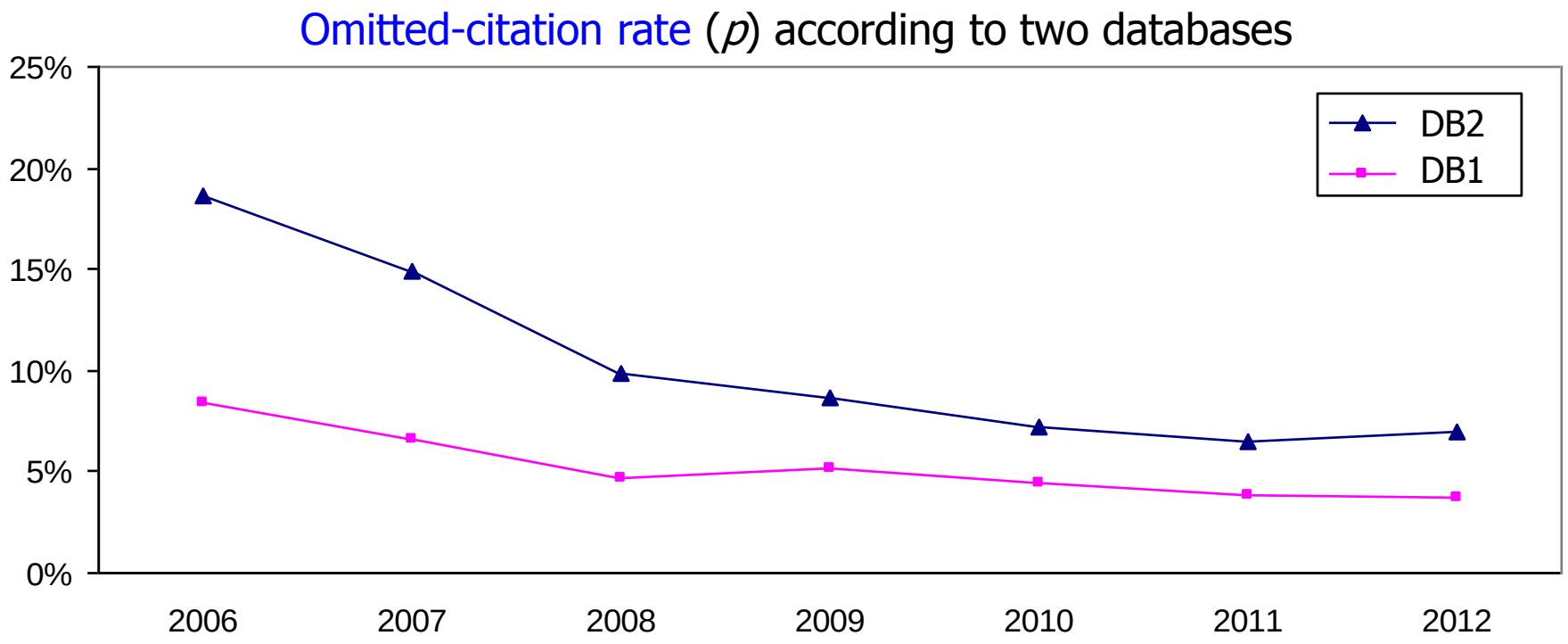
Average citations per paper (e.g., journal impact factor)

# Two practical tools available for our research

- Automated algorithm for estimating the omitted citation rate (*p*) of databases and identifying potential database errors.

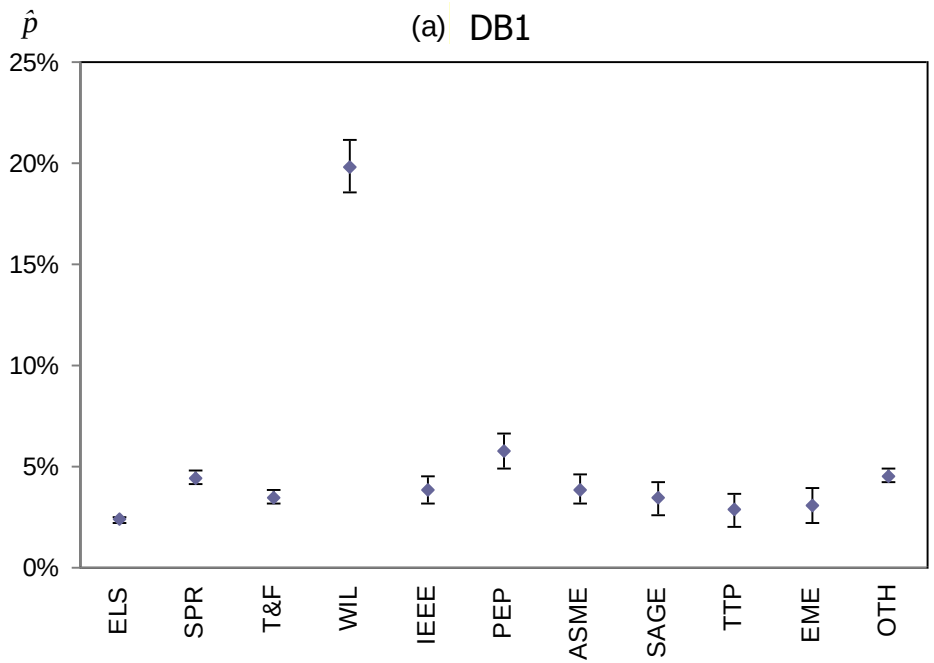
- Statistical model for correcting bibliometric indicators, compensating for omitted citations.

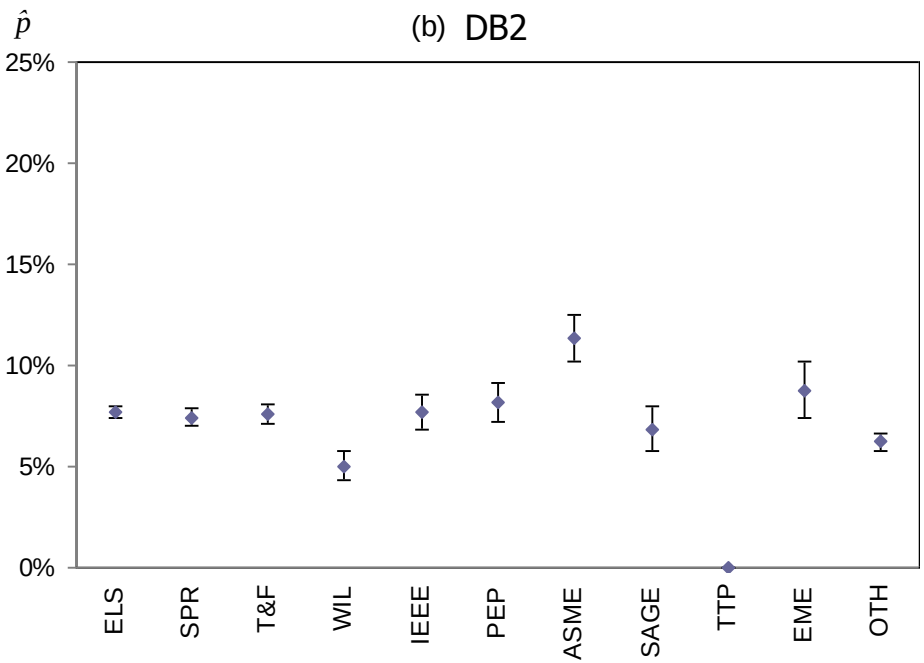
# 1. Gradual reduction in the omitted citation rate




Franceschini, F., Maisano D., Mastrogiacomo L. (2015) Influence of omitted citations on the bibliometric statistics of the major Manufacturing journals. Scientometrics, 103(3): 1083-1122.

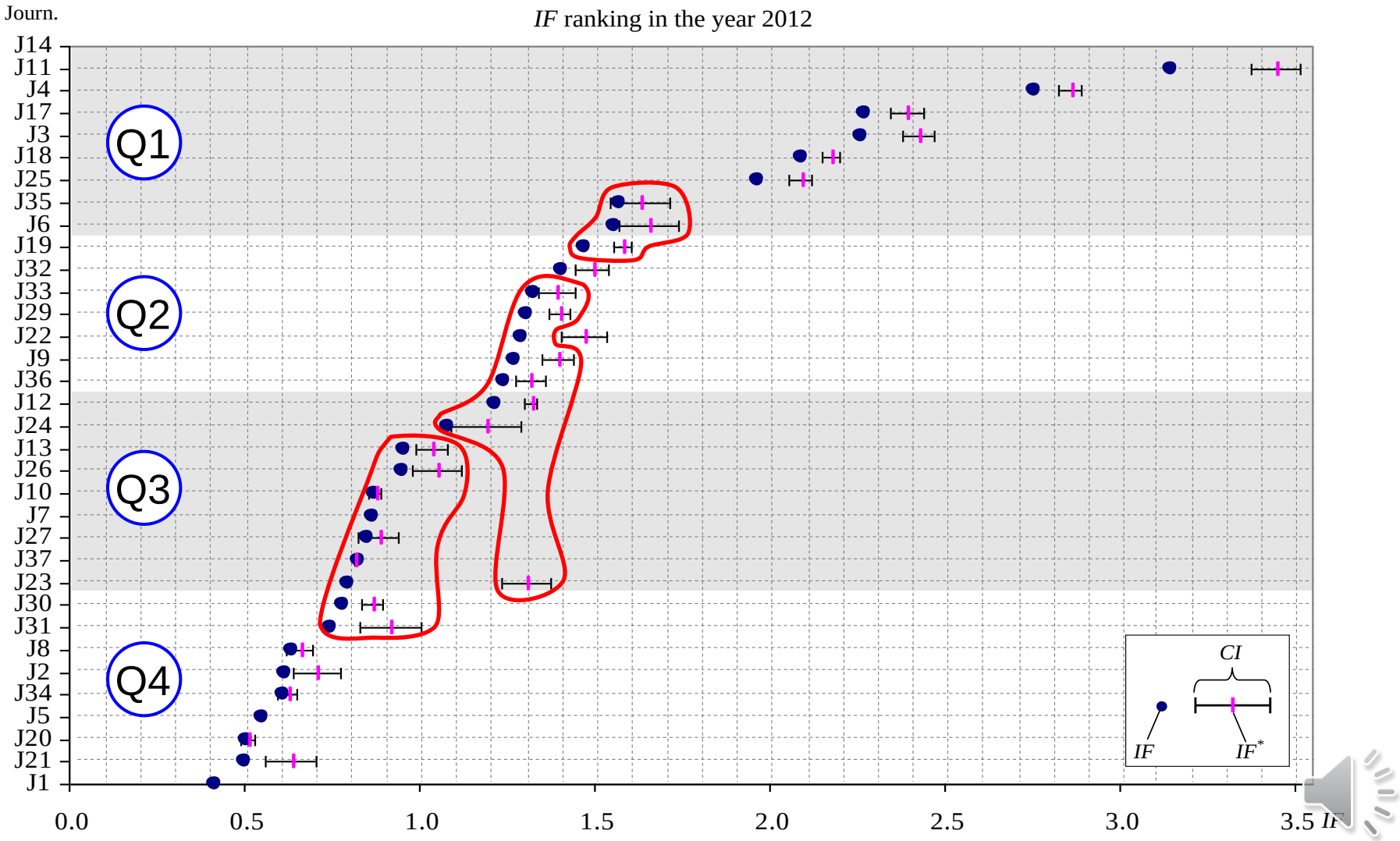
# 2. Effect of editorial styles




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.

# 3. Correction of bibliometric indicators



# 4. Bibliometric errors/horrors

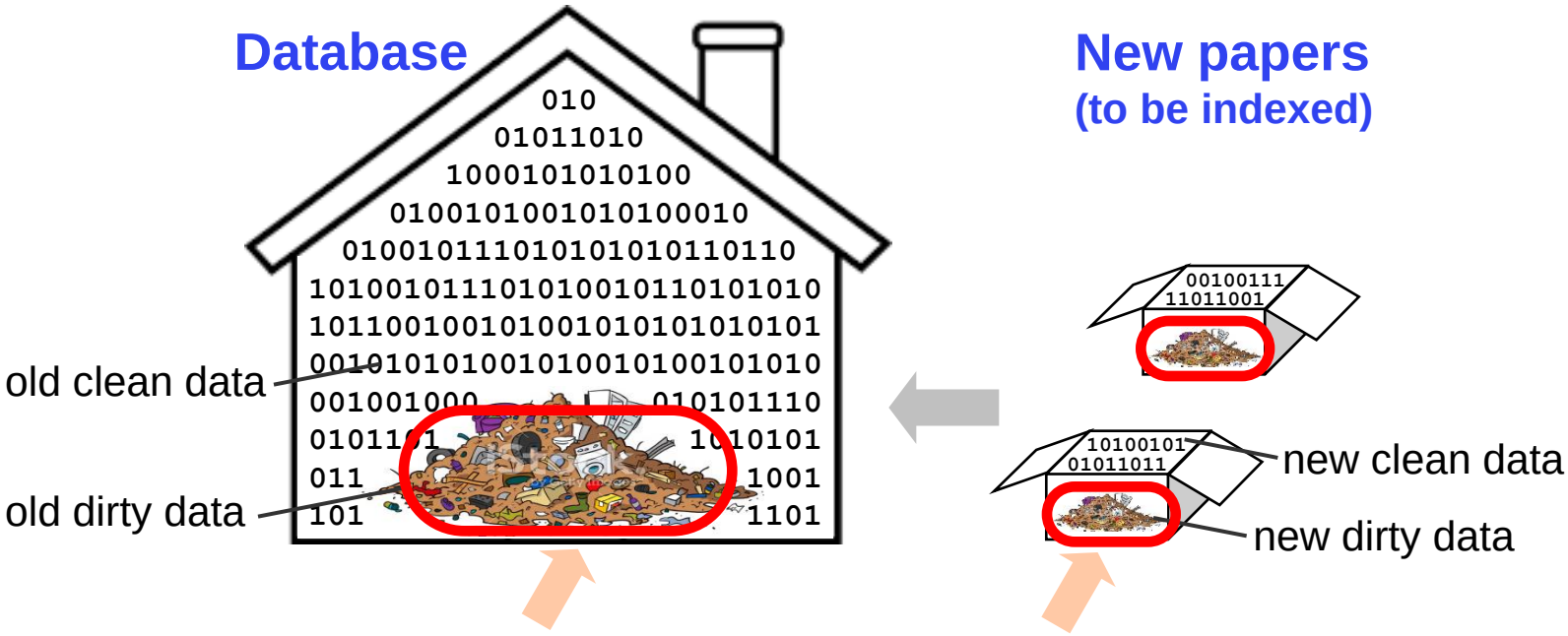
- Loss of citations obtained by Online-First papers.
- Duplication of DOI (Digital Object Identifier) codes.
- "Disintegrated" references.
- "Imaginary" references.

…

Franceschini, F., Maisano D., Mastrogiacomo L. (2015) Errors in DOI indexing by bibliometric databases. Scientometrics, 102(3): 2181-2186.

Franceschini, F., Maisano D., Mastrogiacomo L. (2014) The museum of errors/horrors in Scopus. Journal of Informetrics, 10(1): 174-182.

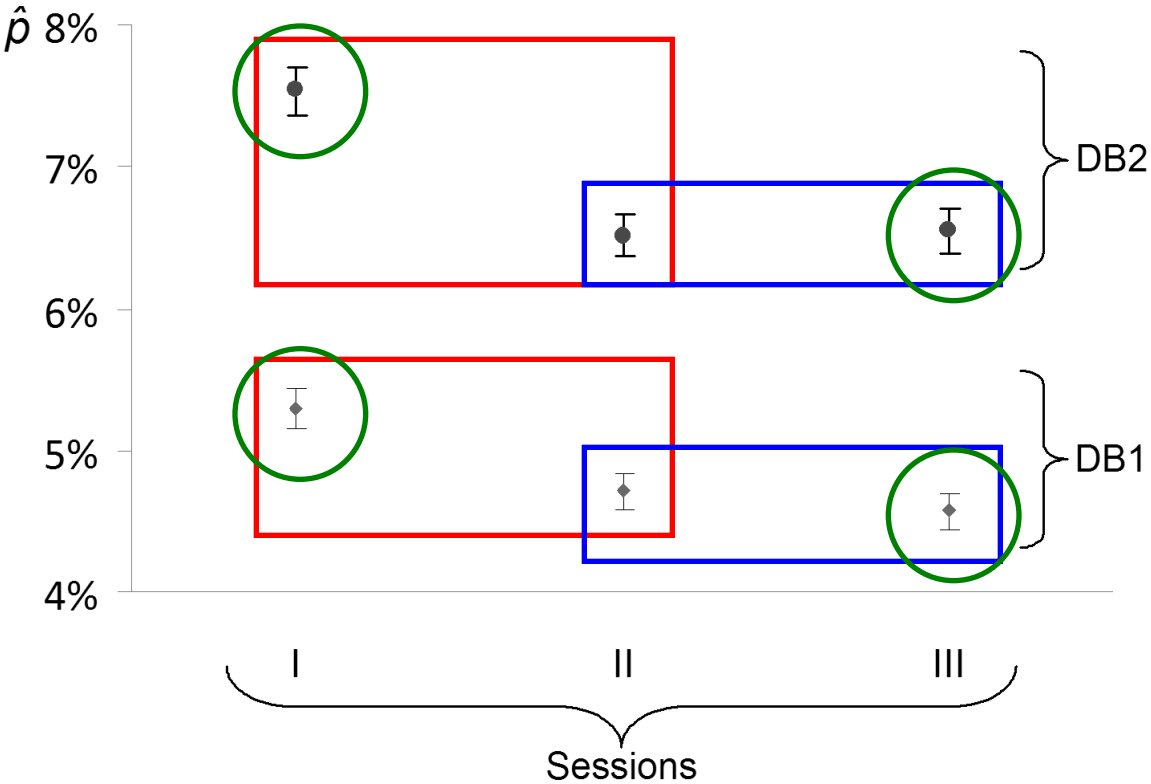
# 5. Improvement by correcting “old” dirty data (1)




1. Reducing the introduction of **new** dirty data,
   i.e., errors concerning new papers to be indexed;

2. Cleaning up **old** dirty data,
   i.e., errors concerning papers/sources already indexed.

# 5. Improvement by correcting “old” dirty data (2)




$$\Delta p = \left( \frac{\hat{p}_{final} - \hat{p}_{initial}}{\hat{p}_{initial}} \right)$$

| | $\Delta p$ | | |
|---|---|---|---|
| | I to II | II to III | I to III |
| DB1 | -11.1% | -2.9% | -13.7% |
| DB2 | -13.5% | 0.4% | -13.1% |

# 6. Classification of database errors (1)

| | | DB1 | DB2 |
|---|---|---|---|
| **Type-A or pre-existing errors** | | (Freq.) | (Freq.) |
| A.1 | Missing/wrong article title | 0.41% | 0.93% |
| A.2 | Errors in the other fields | 0.18% | 1.01% |
| | **Subtotal** | **0.59%** | **1.95%** |
| **Type-B or database mapping errors** | | | |
| B.1 | Errors in the transcription of author name(s) and/or article title | 0.13% | 1.65% |
| B.2 | Incomplete cited-article list | 0.09% | 0.11% |
| B.3 | Omitted cited-article list | 0.08% | 0.14% |
| B.4 | Wrong or missing DOI | 0.09% | 0.14% |
| B.5 | Errors concerning Online-First articles | 0.76% | 0.69% |
| B.6 | Unindexed (citing) articles | 1.30% | 0.16% |
| B.7 | Reasons unknown | 1.07% | 1.61% |
| | **Subtotal** | **3.53%** | **4.51%** |
| | **Total (*p'*)** | **4.12%** | **6.46%** |

# 6. Classification of database errors (2)

## Example of "A.2-Errors in other fields"

**Cited paper ($P_1$):**

Authors: J. Dong, P.M. Ferreira, J.A. Stori
Title: Feed-rate optimization with jerk constraints for generating minimum-time trajectories
Source: International Journal of Machine Tools and Manufacture, 47(12-13): 1941-1955
DOI: 10.1016/j.ijmachtools.2007.03.006

**Citing paper ($P_2$):**

Authors: X. Broquere, D. Sidobre, I. Herrera-Aguilar
Title: Soft motion trajectory planner for service manipulator robot
Source: IEEE/RSJ International Conference on Intelligent Robots and Systems, 2008. IROS 2008.
DOI: 10.1109/IROS.2008.4650608

**Reference to $P_1$ (with inaccurate author names), in the list of $P_2$:**

[8] P. F. Jingyan Dong and J. Stori, "Feed-rate optimization with jerk constraints for generating minimum-time trajectories," *International Journal of Machine Tools and Manufacture*, 2007.

**Reference to $P_1$ (with inaccurate author names), according to Scopus:**

Scopus

7 Kyriakopoulos, Konstantinos J., Saridis, George N.
MINIMUM JERK PATH GENERATION.
(1988) , pp. 364-369. Cited 72 times.
ISBN: 0818608528
POLITO SFX BETA View at Publisher

8 Jingyan Dong, P.F., Stori, J.
Feed-rate optimization with jerk constraints for generating minimum-time trajectories
(2007) *International Journal of Machine Tools and Manufacture*. Cited 2 times.

missing link to $P_1$

## Example of "B.3-Omitted cited-article list"

**(Citing) paper of interest ($P_1$):**

Authors: J. Hong, D. Xu, P. Gong, J. Yu, H. Ma, S. Yao
Title: Covalent-bonded immobilization of enzyme on hydrophilic polymer covering magnetic nanogels
Source: Microporous and Mesoporous Materials, 109(1-3): 470-477
DOI: 10.1016/j.micromeso.2007.05.052

**Original list of ($P_1$):**

**References**

[1] K.M. Koeller, C.H. Wong, Nature 409 (2001) 232.
[2] R. Sharma, Y. Chisti, U.C. Banerjee, Biotechnol. Adv. 19 (2001) 627.
[...]
[37] S. Rauf, A. Ihsan, K. Akhtar, M.A. Ghauri, M. Rahman, M.A. Anwar, A.M. Khalid, J. Biotechnol. 121 (2006) 351.
[38] S. Tembe, M. Karve, S. Inamdar, S. Haram, J. Melo, S.F. D'Souza, Anal. Biochem. 349 (2006) 72.

**Missing list in WoS:**

WEB OF SCIENCE™

Covalent-bonded immobilization of enzyme on hydrophilic polymer covering magnetic nanogels
By: Hong, J (Hong, J.); Xu, D (Xu, D.); Gong, P (Gong, P.); Yu, J (Yu, J.); Ma, H (Ma, H.); Yao, S (Yao, S.)
MICROPOROUS AND MESOPOROUS MATERIALS
Volume: 109 Issue: 1-3 Pages: 470-477
DOI: 10.1016/j.micromeso.2007.05.052
Published: MAR 1 2008

Citation Network
47 Times Cited
0 Cited References
View Citation Map
Create Citation Alert
(data from Web of Science ™ Core Collection)

absence of references

Franceschini, F., Maisano D., Mastrogiacomo L. (2016) Empirical analysis and classification of database errors in Scopus and Web of Science. Journal of Informetrics, 10(4): 933-953.

# Conclusions (1)

## Implications

- Bibliometric errors may distort indicators/statistics concerning the production output of

→ individual scientists (e.g., comparative evaluations);

→ groups of scientists (e.g., evaluations of departments, universities, etc.);

→ scientific journals.

- The systematic application of the automated algorithm would make it possible to address potential errors directly.

# Conclusions (2)

## Limitations

- The concurrent omission of a citing paper by both databases will prevent its detection.
- Cited papers are all confined within the Engineering-Manufacturing field.

## Future research

- Extending the study to scientific papers in other fields, considering a longer time-scale (e.g., 10-15 years).

# Thank you for your attention

domenico.maisano@polito.it

http://orcid.org/0000-0002-8154-4469

http://staff.polito.it/fiorenzo.franceschini/Maisano_Pub.htm