

On the correction of “old” omitted citations by bibliometric databases

Original

On the correction of “old” omitted citations by bibliometric databases / Franceschini, Fiorenzo; Maisano, DOMENICO AUGUSTO FRANCESCO; Mastrogiacomo, Luca. - ELETTRONICO. - (2015), pp. 1200-1207. (15th ISSI (International Society of Scientometrics and Informetrics Conference) 2015 Istanbul, Turkey 29 June - 4 July 2015).

Availability:

This version is available at: 11583/2614476 since: 2015-07-06T14:59:39Z

Publisher:

Boaziçi University Printhouse

Published

DOI:

Terms of use:

This article is made available under terms and conditions as specified in the corresponding bibliographic description in the repository

Publisher copyright

(Article begins on next page)

ISSI
2015

15th INTERNATIONAL CONFERENCE ON SCIENTOMETRICS & INFORMETRICS

29 June - 4 July, 2015
BOĞAZIÇI UNIVERSITY • ISTANBUL-TURKEY
www.issi2015.org



PROCEEDINGS OF ISSI 2015



PROCEEDINGS OF ISSI 2015 ISTANBUL

**15th International Society of
Scientometrics and Informetrics Conference**

**Istanbul, Turkey
29th June to 4th July 2015**

Editors

Albert Ali Salah, Yaşar Tonta, Alkım Almıla Akdağ Salah,
Cassidy Sugimoto, Umut Al

Partners

Boğaziçi University, Turkey
Hacettepe University, Turkey
TÜBİTAK ULAKBİM, Turkey

Sponsors

Thomson Reuters
Springer
EBSCO Information Services, USA
Emerald
Elsevier B.V.

Boğaziçi University Cataloging –in-Publication Data

Proceedings of ISSI 2015 Istanbul: 15th International Society of Scientometrics and Informetrics Conference, Istanbul, Turkey, 29 June to 3 July, 2015 / editors Albert Ali Salah, Yaşar Tonta, Alkım Almıla Akdağ Salah, Cassidy Sugimoto, Umut Al.

1275 p.; 29 cm.

ISBN 978-975-518-381-7

ISSN 2175-1935

1. Bibliometrics - Congresses. 2. Information science - Congresses. 3. Communication in science - Congresses. 4. Scientific literature - Congresses.
Z669.8jb.I58

Printed at the Boğaziçi University Printhouse

First Printing: June 2015

Boğaziçi Üniversitesi, ETA B Blok, Zemin Kat, Kuzey Kampüs, İstanbul / TÜRKİYE

Tel ve Fax: (90) 212 359 44 06

ORGANISATION AND COMMITTEES

Conference Chairs

Albert Ali Salah
Yaşar Tonta
M. Mirat Satoğlu

Programme Chairs

Alkım Almila Akdağ Salah
Cassidy Sugimoto
Umut Al

Doctoral Consortium Chairs

Andrea Scharnhorst
Judit Bar-Ilan

Workshops & Tutorials Chair

Caroline Wagner

Local Organization Chair

Heysem Kaya

Scientific Program Committee

Giovanni Abramo
Isidro Aguillo
Isola Ajiferuke
Alkım Almila Akdağ Salah
Umut Al
Jens-Peter Andersen
Eric Archambault
Clément Arsenault
Joaquin Azagra-Caro
Judit Bar-Ilan
Aparna Basu
Sada Bihari-Sahu
Maria Bordons
Lutz Bornmann
Hamid Bouabid
Kevin W. Boyack
Guillaume Cabanac
Juan Miguel Campanario
Chaomei Chen
Andrea D'Angelo
Hans-Dieter Daniel
Cinzia Daraio
Hamid Darvish
Koenraad Debackere
Gernot Deinzer
Brad Demarest
Swapan Deoghuria

Fereshteh Didegah
Ying Ding
Güleda Doğan
Tim Engels
Claire François
Jonathan Furner
Antonio Garcia
Aldo Geuna
Wolfgang Glänzel
Alicia Gomez
Isabel Gomez
Juan Gorraiz
Philippe Gorry
Abdullah Gök
Raf Guns
Nabi Hasan
Stefanie Haustein
Sybille Hinze
Michael Hofer
Marianne Hörlesberger
Zhigang Hu
Peter Ingwersen
Siladitya Jana
Evaristo Jiménez-Contreras
Milos Jovanovic
Yuya Kajikawa
Hiltrun Kretschmer

ORGANISATION AND COMMITTEES

J P S Kumaravel
Benedetto Lepori
Jacqueline Leta
Jonathan Levitt
Loet Leydesdorff
Liming Liang
Judith Licea
Junwan Liu
Yuxian Liu
Szu-Chia Lo
Carmen Lopez Illesca
Bob Losee
Terttu Luukkonen
Marc Luwel
Domenico Maisano
Wolfgang Mayer
Kate McCain
Eustache Megnigbeto
Lokman Meho
Raul Mendez-Vasquez
Alexis Michel Mugabushaka
Ulle Must
Anton Nederhof
Ed Noyons
Michael Ochsner
Carlos Olmeda-Gómez
José Luis Ortega
Maria Antonia Ovalle-Perandonos
Adèle Paul-Hus
Antonio Perianes-Rodríguez
Bluma C. Peritz
Fernanda Peset
Anastassios Pouris
Ismael Rafols
Emanuela Reale
John Rigby
Nicolas Robinson-Garcia
Ivana Roche
Jürgen Roth
Ronald Rousseau
Santanu Roy
Jane Russell
Bibhuti Sahoo
Albert Ali Salah
Ulf Sandström
Elias Sanz
Andrea Scharnhorst
Edgar Schiebel

Christian Schloegl
Jesper Schneider
Torben Schubert
Robert Shelton
Gunnar Sivertsen
Stig Slipersæter
Henry Small
Andreas Strotmann
Cassidy Sugimoto
Yuan Sun
Zehra Taşkın
Mike Thelwall
Bart Thijs
Yaşar Tonta
Andrew Tsou
Saeed Ul Hassan
Peter van den Besselaar
Nees Jan Van Eck
Thed Van Leeuwen
Bart Van Looy
Anthony Van Raan
Benjamin Vargas-Quesada
Liwen Vaughan
Peter Vinkler
Cathelijn Waaijer
Lili Wang
Xianwen Wang
Jos Winnink
Matthias Winterhager
Dietmar Wolfram
Paul Wouters
Qiang Wu
Yishan Wu
Erjia Yan
Dangzhi Zhao
Michel Zitt
Alesia Zuccala

CHAIRS' WELCOME

The 15th International Society of Scientometrics and Informetrics Conference took place at Boğaziçi University in Istanbul, from June 29 to July 4, 2015. The Conference was jointly organised by Boğaziçi University, Hacettepe University, and the TÜBİTAK ULAKBİM (Turkish Academic Network and Information Center – The Scientific and Technological Research Council of Turkey) under the auspices of ISSI – the International Society for Scientometrics and Informetrics.

The ISSI biennial conference is the premier international forum for scientists, research managers, authorities and information professionals to discuss the current status and progress in informetric and scientometric theories, concepts, tools, platforms, and indicators. In addition to theoretical and quantitative focus of the conference, the participants had the opportunity to discuss practical, cross-cultural, and multi-disciplinary aspects of information and library science, R&D-management, and science ethics, among other related topics.

The focus theme of ISSI2015 was “**the future of scientometrics**”. Scientometrics and informetrics together represent a broad field with a rich history. Scientometrics has been responsible for creating tools for research assessment and evaluation, as well as for use in charting the flow of scientific ideas and people. Today, with the advancements of computing power, technology, and database management systems, the impact of scientometrics has become ubiquitous for scientists and science policy makers. However, the high diffusion of scientometric and informetric research has also brought a new wave of criticism and concern, as people grapple with issues of goal displacement and inappropriate use of indicators. The question facing the field is how best to move forward given the computational opportunities and the sociological concerns. Therefore, the goal of ISSI2015 was to highlight the best research in this field and to bring together scholars and practitioners in the area to discuss new research directions, methods, and theories, and to reflect upon the history of scientometrics and its implications.

The keynote given by Loet Leydesdorff demonstrated the potential of thinking of science as a complex institution. By building on the Triple Helix Model of University-Industry-Government relations, Dr. Leydesdorff showed that innovation systems can provide institutional mediation between knowledge production, wealth generation, and governance.

The second keynote, by Kevin Boyack, directly answered the challenge of the focus theme of ISSI2015, and proposed several opportunities to expand the field of scientometrics. Dr. Boyack called for increasing attention to funding, workforce, data and instrumentation, research objects, and innovation.

The conference included four special sessions on a range of topics, including performance indicators, algorithms for topic detection, empirical evaluation of education, research and innovation, and how scientometrics can be used to improve and inform university rankings. These special sessions included poster presentations, panel discussions, invited speakers, and public debates.

The increasing number of open-source software for scientometrics presents great opportunities for researchers. Four tutorials, organized on the first day of the conference, aimed to introduce a number of tools in depth: open source data analysis and visualization tools, citation exploration software, measurement of scholarly impact, and on social network analysis with the popular R software.

The Doctoral Forum, organized by Andrea Scharnhorst and Judit Bar-Ilan, is a meeting of senior researchers and selected doctoral students for presenting and discussing research projects and an

excellent way for students of getting valuable feedback, along with strong networking opportunities. This is the sixth ISSI Doctoral Forum and we are extremely happy about the interest it continues to receive from the community. Additionally, the prestigious Eugene Garfield Doctoral Dissertation Scholarship is given by the Eugene Garfield Foundation.

During the Conference, the Derek de Solla Price Award of the International Journal Scientometrics was given to Mike Thelwall, Professor of Information Science at the University of Wolverhampton (UK), in a special session organized for this purpose. This award recognizes excellence through outstanding, sustained career achievements in the field of quantitative studies of science and their applications.

The satellite workshops of the conference reflected the diversity of the field. In **“Mining Scientific Papers: Computational Linguistics and Bibliometrics”**, researchers in bibliometrics and computational linguistics were brought together to study the ways bibliometrics can benefit from large-scale text analytics and sense mining of scientific papers, thus exploring the interdisciplinarity of Bibliometrics and Natural Language Processing. The workshop on **“Grand challenges in data integration for research and innovation policy”** dealt with problems of big, open and linked data. The **“Forecasting science: Models of science and technology dynamics for innovation policy”** workshop discussed methodology for predicting the circumstances leading to scientific or technological innovation. **“Workshop on Bibliometrics Education”** brought together educational institutions, employers, professional societies, and Bibliometrics researchers and professionals to tackle this problem. Finally, **“Google Scholar and related products”** was a highly interactive workshop on the benefits and limitations of some of the most important citation tools.

All contributions for the conference were evaluated by at least two reviewers of the Scientific Program Committee. The papers that required additional reviews were discussed by the Program Chairs before a decision was reached. From 228 full and research in progress paper submissions, 123 papers were accepted for publication (54 percent acceptance rate). 82 of these papers were full papers, and 41 were research in progress. There was a large number of paper submissions on social media, technology transfer, science policy and research assessment. From 123 poster and ignite talk submissions, 68 posters and 13 ignite talks were accepted (66 percent). The ignite talks were to increase discussion of underrepresented topics and novel ideas. Because of the large number of papers, and to allow proper discussion for each paper, four parallel sessions were implemented. Several poster sessions were organized, each containing a relatively manageable number of posters. The conference brought together researchers from 42 countries and the works of 458 researchers were presented.

We thank all our contributors for their submissions, the members of the Organizing Committee for their work, the Scientific Program Committee for their reviewing effort, the ISSI board for their trust and guidance, the Rectorate and the Faculty of Engineering of Boğaziçi University for their constant assistance and support, as well as the sponsors for their generous financial contributions. We particularly thank Metin Tunç (Thomson Reuters), Elif Gürses (formerly of TÜBİTAK ULAKBİM), Juan Gorraiz (Universitat Wien), Figen Atalan (Boğaziçi University), Orçun Madran (Hacettepe University) and Büşra Şahin (DEKON Congress & Tourism) for their help in organizing ISSI2015.

Albert Ali Salah, Yaşar Tonta, Mirat Satoğlu, Alkım Almıla Akdağ Salah, Cassidy Sugimoto, Umut Al

TABLE OF CONTENTS

ALTMETRICS/WEBOMETRICS	PAGE
Who Publishes, Reads, and Cites Papers? An Analysis of Country Information <i>Robin Haunschild, Moritz Stefaner, and Lutz Bornmann</i>	4
Do Mendeley Readership Counts Help to Filter Highly Cited WoS Publications better than Average Citation Impact of Journals (JCS)? <i>Zohreh Zahedi, Rodrigo Costas and Paul Wouters</i>	16
Influence of Study Type on Twitter Activity for Medical Research Papers <i>Jens Peter Andersen and Stefanie Haustein</i>	26
Is There a Gender Gap in Social Media Metrics? <i>Adèle Paul-Hus, Cassidy R. Sugimoto, Stefanie Haustein and Vincent Larivière</i>	37
PubMed and ArXiv vs. Gold Open Access: Citation, Mendeley, and Twitter Uptake of Academic Articles of Iran <i>Ashraf Maleki</i>	46
Alternative Metrics for Book Impact Assessment: Can Choice Reviews be a Useful Source? <i>Kayvan Kousha and Mike Thelwall</i>	59
A Longitudinal Analysis of Search Engine Index Size <i>Antal van den Bosch, Toine Bogers and Maurice de Kunder</i>	71
Online Attention of Universities in Finland: Are the Bigger Universities Bigger Online too? <i>Kim Holmberg</i>	83
Ranking Journals Using Altmetrics <i>Tamar V. Loach and Tim S. Evans</i>	89
Who Tweets about Science? <i>Andrew Tsou, Tim Bowman, Ali Ghazinejad, and Cassidy Sugimoto</i>	95
Classifying Altmetrics by Level of Impact <i>Kim Holmberg</i>	101
Characterizing In-Text Citations Using N-Gram Distributions <i>Marc Bertin and Iana Atanassova</i>	103
Can Book Reviews be Used to Evaluate Books' Influence? <i>Qingqing Zhou and Chengzhi Zhang</i>	105
Adapting Sentiment Analysis for Tweets Linking to Scientific Papers <i>Natalie Friedrich, Timothy D. Bowman, Wolfgang G. Stock and Stefanie Haustein</i>	107
Mendeley Readership Impact of Academic Articles of Iran <i>Ashraf Maleki</i>	109
Does the Global South Have Altmetrics? Analyzing a Brazilian LIS Journal <i>Ronaldo F. Araújo, Tiago R. M. Murakami, Jan L. de Lara and Sibebe Fausto</i>	111
Tweet or Publish: A Comparison of 395 Professors on Twitter <i>Timothy D. Bowman</i>	113
Stratifying Altmetrics Indicators Based on Impact Generation Model <i>Qiu Junping and Yu Houqiang</i>	115

CITATION AND COCITATION ANALYSIS	PAGE
Citation Type Analysis for Social Science Literature in Taiwan <i>Ming-yueh Tsay</i>	117
University Citation Distributions <i>Antonio Perianes-Rodriguez and Javier Ruiz-Castillo</i>	129
Exploration of the Bibliometric Coordinates for the Field of 'Geography' <i>Juan Gorraiz and Christian Gumpenberger</i>	139
The Most-Cited Articles of the 21st Century <i>Elias Sanz-Casado, Carlos García-Zorita and Ronald Rousseau</i>	150
An International Comparison of the Citation Impact of Chinese Journals with Priority Funding <i>Ping Zhou and Loet Leydesdorff</i>	160
Research Data Explored: Citations versus Altmetrics <i>Isabella Peters, Peter Kraker, Elisabeth Lex, Christian Gumpenberger and Juan Gorraiz</i>	172
Stopped Sum Models for Citation Data <i>Wan Jing Low, Paul Wilson and Mike Thelwall</i>	184
Differences in Received Citations over Time and Across Fields in China <i>Siluo Yang, Junping Qiu, Jinda Ding and Houqiang Yu</i>	195
The Rise in Co-authorship in the Social Sciences (1980-2013) <i>Dorte Henriksen</i>	209
The Recurrence of Citations within a Scientific Article <i>Zhigang Hu, Chaomei Chen and Zeyuan Liu</i>	221
Do Authors with Stronger Bibliographic Coupling Ties Cite Each Other More Often? <i>Ali Gazni and Fereshteh Didegah</i>	230
The Research of Paper Influence Based on Citation Context - A Case Study of the Nobel Prize Winner's Paper <i>Shengbo Liu, Kun Ding, Bo Wang, Delong Tang and Zhao Qu</i>	241
Time to First Citation Estimation in the Presence of Additional Information <i>Tina Nane</i>	249
Author Relationship Mining based on Tripartite Citation Analysis <i>Feifei Wang, Junwan Liu and Siluo Yang</i>	261
Charles Dotter and the Birth of Interventional Radiology: A "Sleeping-Beauty" with a Restless Sleep <i>Philippe Gorry and Pascal Ragouet</i>	266
Citation Distribution of Individual Scientist: Approximations of Stretch Exponential Distribution with Power Law Tails <i>Ol. S. Garanina and Michael Yu. Romanovsky</i>	272
Influence of International Collaboration on the Research Impact of Young Universities <i>Khiam Aik Khor and Ligen G. Yu</i>	278
Which Collaborating Countries Give to Turkey the Largest Amount of Citation? <i>Bárbara S. Lancho Barrantes</i>	280
Do We Need Global and Local Knowledge of the Citation Network? <i>Sophia. R. Goldberg, Hannah Anthony and Tim S. Evans</i>	282

Citation Analysis as an Auxiliary Decision-Making Tool in Library Collection Development	284
<i>Iva Vrkić</i>	
Is Paper Uncitedness a Function of the Alphabet?	286
<i>Clément Arsenault and Vincent Larivière</i>	
Relative Productivity Drivers of Economists: A Probit/Logit Approach for Six European Countries	288
<i>Stelios Katranidis and Theodore Panagiotidis</i>	
Do First-Articles in a Journal Issue Get More Cited?	290
<i>Tian Ruiqiang, Yao Changqing, Pan Yuntao, Wu Yishan, Su Cheng and Yuan Junpeng</i>	
Proquest Dissertation Analysis	292
<i>Kishor Patel, Sergio Govoni, Ashwini Athavale, Robert P. Light and Katy Börner</i>	

INDICATORS	PAGE
An Alternative to Field-Normalization in the Aggregation of Heterogeneous Scientific Fields	294
<i>Antonio Perianes-Rodriguez and Javier Ruiz-Castillo</i>	
Correlating Libcitations and Citations in the Humanities with WorldCat and Scopus Data	305
<i>Alesia Zuccala and Howard D. White</i>	
A Vector for Measuring Obsolescence of Scientific Articles	317
<i>Jianjun Sun, Chao Min and Jiang Li</i>	
Field-Normalized Citation Impact Indicators and the Choice of an Appropriate Counting Method	328
<i>Ludo Waltman and Nees Jan van Eck</i>	
Forecasting Technology Emergence from Metadata and Language of Scientific Publications and Patents	340
<i>Olga Babko-Malaya, Andy Seidel, Daniel Hunter, Jason HandUber, Michelle Torrelli and Fotis Barlos</i>	
Understanding Relationship between Scholars' Breadth of Research and Scientific Impact	353
<i>Shiyan Yan and Carl Lagoze</i>	
Transforming the Heterogeneity of Subject Categories into a Stability Interval of the MNCS	365
<i>Marion Schmidt and Daniel Sirtes</i>	
Measuring Interdisciplinarity of a Given Body of Research	372
<i>Qi Wang</i>	
How often are Patients Interviewed in Health Research? An Informetric Approach	384
<i>Jonathan M. Levitt and Mike Thelwall</i>	
Normalized International Collaboration Score: A Novel Indicator for Measuring International Co-Authorship	390
<i>Adam Finch, Kumara Henadeera and Marcus Nicol</i>	
Bibliometric Indicators of Interdisciplinarity Exploring New Class of Diversity Measures	397
<i>Alexis-Michel Mugabushaka, Anthi Kyriakou and Theo Papazoglou</i>	

Modeling Time-dependent and -independent Indicators to Facilitate Identification of Breakthrough Research Papers	403
<i>Holly N. Wolcott, Matthew J. Fouch, Elizabeth Hsu, Catherine Bernaciak, James Corrigan and Duane Williams</i>	
Dimensions of The Author Citation Potential	409
<i>Pablo Dorta-González, María-Isabel Dorta-González and Rafael Suárez-Vega</i>	
Scholarly Book Publishers in Spain: Relationship between Size, Price, Specialization and Prestige	411
<i>Jorge Mañana-Rodríguez and Elea Giménez Toled</i>	
Bootstrapping to Evaluate Accuracy of Citation-Based Journal Indicators	413
<i>Jens Peter Andersen and Stefanie Haustein</i>	
The Lack of Stability of the Impact Factor of the Mathematical Journals	415
<i>Antonia Ferrer-Sapena, Enrique A. Sánchez-Pérez, Fernanda Peset, Luis-Millán González and Rafael Aleixandre-Benavent</i>	
Using Bibliometrics to Measure the Impact of Cancer Research on Health Service and Patient Care: Selecting and Testing Four Indicators	417
<i>Frédérique Thonon, Mahasti Saghatchian, Rym Boulkedid and Corinne Alberti</i>	
A New Scale for Rating Scientific Publications	419
<i>Răzvan Valentin Florian</i>	
Analysis of the Factors Affecting Interdisciplinarity of Research in Library and Information Science	421
<i>Chizuko Takei, Fuyuki Yoshikane and Hiroshi Itsumura</i>	
An Analysis of Scientific Publications from Serbia: The Case of Computer Science	423
<i>Miloš Pavković and Jelica Protić</i>	

SCIENCE POLICY AND RESEARCH ASSESSMENT	PAGE
A Computer System for Automatic Evaluation of Researchers' Performance	425
<i>Ashkan Ebadi and Andrea Schiffauerova</i>	
Grading Countries/Territories Using DEA Frontiers	436
<i>Guo-liang Yang, Per Ahlgren, Li-ying Yang, Ronald Rousseau and Jie-lan Ding</i>	
Continuous, Dynamic and Comprehensive Article-Level Evaluation of Scientific Literature	448
<i>Xianwen Wang, Zhichao Fang and Yang Yang</i>	
Interdisciplinarity and Impact: Distinct Effects of Variety, Balance, and Disparity	460
<i>Jian Wang, Bart Thijs and Wolfgang Glänzel</i>	
The Evaluation of Scholarly Books as a Research Output. Current Developments in Europe	469
<i>Elea Giménez-Toledo, Jorge Mañana-Rodríguez, Tim Engels, Peter Ingwersen, Janne Pölonen, Gunnar Sivertsen, Frederik Verleysen and Alesia Zuccala</i>	
Publications or Citations – Does it Matter? Beneficiaries in Two Different Versions of a National Bibliometric Performance Model, an Existing Publication-based and a Suggested Citation-based Model	477
<i>Jesper W. Schneider</i>	
The Effect of Having a Research Chair on Scientists' Productivity	489
<i>Seyed Reza Mirnezami and Catherine Beaudry</i>	

Drivers of Higher Education Institutions' Visibility: A Study of UK HEIs Social Media Use vs. Organizational Characteristics	502
<i>Julie M. Birkholz, Marco Seeber and Kim Holmberg</i>	
A Computing Environment to Support Repeatable Scientific Big Data Experimentation of World-Wide Scientific Literature	514
<i>Bob G. Schlicher, James J. Kulesz, Robert K. Abercrombie, and Kara L. Kruse</i>	
Is Italy a Highly Efficient Country in Science?	525
<i>Aparna Basu</i>	
Performance Assessment of Public-Funded R&D Organizations	537
<i>Debnirmalya Gangopadhyay, Santanu Roy and Jay Mitra</i>	
Outlining the Scientific Activity Profile of Researchers in the Social Sciences and Humanities in Spain: The Case of CSIC	548
<i>Adrián A. Díaz-Faes, María Bordons, Thed van Leeuwen and M^a Purificación Galindo</i>	
A Bibliometric Assessment of ASEAN's Output, Influence and Collaboration in Plant Biotechnology	554
<i>Jane G. Payumo and Taurean C. Sutton</i>	
Science and Technology Indicators In & For the Peripheries. A Research Agenda	560
<i>Ismael Rafols, Jordi Molas-Gallart and Richard Woolley</i>	
Patterns of Internationalization and Criteria for Research Assessment in the Social Sciences and Humanities	565
<i>Gunnar Sivertsen</i>	
Looking for a Better Shape: Societal Demand and Scientific Research Supply on Obesity	571
<i>Lorenzo Cassi, Ismael Rafols, Pierre Sautier and Elisabeth de Turckheim</i>	
Does Quantity Make a Difference?	577
<i>Peter van den Besselaar and Ulf Sandström</i>	
On Decreasing Returns to Scale in Research Funding	584
<i>Philippe Mongeon, Christine Brodeur, Catherine Beaudry and Vincent Larivière</i>	
How Many is too Many? On the Relationship between Output and Impact in Research	590
<i>Vincent Larivière and Rodrigo Costas</i>	
Research Assessment and Bibliometrics: Bringing Quality Back In	596
<i>Michael Ochsner and Sven E. Hug</i>	
Under-Reporting Research Relevant to Local Needs in The Global South. Database Biases in the Representation of Knowledge on Rice	598
<i>Ismael Rafols, Tommaso Ciarli and Diego Chavarro</i>	
Network DEA Approach for Measuring the Efficiency of University- Industry Collaboration Innovation: Evidence from China	600
<i>Yu Yu , Qinfen Shi and Jie Wu</i>	
Promotions, Tenures and Publication Behaviours: Serbian Example	602
<i>Dejan Pajić and Tanja Jevremov</i>	
The Serbian Citation Index: Contest and Collapse	604
<i>Dejan Pajić</i>	
Selecting Researchers with a Not Very Long Career - The Role of Bibliometrics	606
<i>Elizabeth S. Vieira and José A. N.F. Gomes</i>	

Differences By Gender and Role in PhD Theses on Sociology in Spain	608
<i>Lourdes Castelló Cogollos, Rafael Aleixandre Benavent and Rafael Castelló Cogollos</i>	
The Trends to Multi-Authorship and International Collaborative in Ecology Papers	610
<i>João Carlos Nabout, Marcos Aurélio de Amorim Gomes , Karine Borges Machado , Barbara da Silva Rocha , Meirielle Euripa Pádua de Moura , Raquel Menestrino Ribeiro , Lorraine dos Santos Rocha, José Alexandre Felizola Diniz-Filho and Ramiro Logares</i>	
A Bootstrapping Method to Assess Software Impact in Full-Text Papers	612
<i>Erjia Yan and Xuelian Pan</i>	
Article and Journal-Level Metrics in Massive Research Evaluation Exercises: The Italian Case	614
<i>Marco Malgarini, Carmela Anna Nappi and Roberto Torrini</i>	
Accounting For Compositional Effects in Measuring Inter-Country Research Productivity Differences: The Case of Economics Departments in Four European Countries	616
<i>Giannis Karagiannis and Stelios Katranidis</i>	
Metrics 2.0 for Science	618
<i>Isidro F. Aguillo</i>	
Evolution Of Research Assessment In Lithuania 2005 – 2015	620
<i>Saulius Maskeliūnas, Ulf Sandström and Eleonora Dagienė</i>	
Research-driven Classification and Ranking in Higher Education: An Empirical Appraisal of a Romanian Policy Experience	622
<i>Gabriel-Alexandru Vîiu, Mihai Păunescu, and Adrian Miroiu</i>	
Looking beyond the Italian VQR 2004-2010: Improving the Bibliometric Evaluation of Research	634
<i>Alberto Anfossi, Alberto Cioffi and Filippo Costa</i>	
High Fluctuations of THES-Ranking Results in Lower Scoring Universities	640
<i>Johannes Sorz, Martin Fieder, Bernard Wallner and Horst Seidler</i>	
The Vicious Circle of Evaluation Transparency – An Ignition Paper	646
<i>Miloš Jovanović</i>	
Influence of the Research-Oriented President’s Competency on Research Performance in University of China – Based on the Results of Empirical Research	648
<i>Li Gu, Liqiang Ren, Kun Ding and Wei Hu</i>	
Medical Literature Imprinting by Pharma Ghost Writing: A Scientometric Evaluation	650
<i>Philippe Gorry</i>	
Are Scientists Really Publishing More?	652
<i>Daniele Fanelli and Vincent Larivière</i>	
COUNTRY LEVEL STUDIES AND PATENT ANALYSIS	
Tapping into Scientific Knowledge Flows via Semantic Links	654
<i>Saeed-Ul Hassan and Peter Haddawy</i>	
Causal Connections between Scientometric Indicators: Which Ones Best Explain High-Technology Manufacturing Outputs?	662
<i>Robert D. Shelton, Tarek R. Fadel and Patricia Foland</i>	

Scientific Production in Brazilian Research Institutes: Do Institutional Context, Background Characteristics and Academic Tasks Contribute to Gender Differences?	673
<i>Gilda Olinto and Jacqueline Leta</i>	
Comparing the Disciplinary Profiles of National and Regional Research Systems by Extensive and Intensive Measures	684
<i>Irene Bongioanni, Cinzia Daraio, Henk F. Moed and Giancarlo Ruocco</i>	
New Research Performance Evaluation Development and Journal Level Indices at Meso Level	697
<i>Muzammil Tahira, Rose Alinda Alias, Aryati Bakri and A. Abrizah</i>	
Factors Influencing Research Collaboration in LIS Schools in South Africa	707
<i>Jan Resenga Maluleka, Omwoyo Bosire Onyancha and Isola Ajiferuke</i>	
The Diffusion of Nanotechnology Knowledge in Turkey	720
<i>Hamid Darvish and Yaşar Tonta</i>	
The Network Structure of Nanotechnology Research Output of Turkey: A Co-authorship and Co-word Analysis Study	732
<i>Hamid Darvish and Yaşar Tonta</i>	
Analysis of the Spatial Dynamics of Intra- v.s. Inter-Research Collaborations across Countries	744
<i>Lili Wang and Mario Coccia</i>	
Nanotechnology Research in Post-Soviet Russia: Science System Path-Dependencies and their Influences	755
<i>Maria Karaulova, Oliver Shackleton, Abdullah Gök and Philip Shapira</i>	
Support Programs to Increase the Number of Scientific Publications Using Bibliometric Measures: The Turkish Case	767
<i>Yaşar Tonta</i>	
What's Special about Book Editors? A Bibliometric Comparison of Book Editors and other Flemish Researchers in the Social Sciences and Humanities	778
<i>Truyken L.B. Ossenkop and Mike Thelwall</i>	
Scientific Cooperation in the Republics of Former Yugoslavia Before, During and After the Yugoslav Wars	784
<i>Dragan Ivanović, Miloš M. Jovanović and Frank Fritsche</i>	
The Brazilian National Impact: Movement of Journals Between Bradford Zones of Production and Consumption	790
<i>Rogério Mugnaini and Luciano A. Digiampietri</i>	
Sustained Collaboration Between Researchers in Mexico and France in the Field of Chemistry	796
<i>Jane M. Russell, Shirley Ainsworth and Jesús Omar Arriaga-Pérez</i>	
Innovation and Economic Growth: Delineating the Impact of Large and Small Innovators in European Manufacturing	802
<i>Jan-Bart Vervenne and Bart Van Looy</i>	
Chemistry Research in India: A Bright Future Ahead	808
<i>Swapan Deoghuria, Gayatri Paul and Satyabrata Roy</i>	
Main Institutional Sectors in the Publication Landscape of Spain: The Role of Non-Profit Entities	810
<i>Borja González Albo, Javier Aparicio, Luz Moreno-Solano and María Bordons</i>	

Reform of Russian Science as a Reason for Scientometrics Research Growth	812
<i>Andrey Guskov</i>	
Leadership Among the Leaders of The Brazilian Research Groups in Marine Biotechnology	814
<i>Sibele Fausto and Jesús P. Mena-Chalco</i>	
An Empirical Study on Utilizing Pre-grant Publications in Patent Classification Analysis	816
<i>Chung-Huei Kuan and Chan-Yi Lin</i>	
The New Development Trend of Chinese-funded Banks and Internet Financial Enterprises from Patent Perspective	826
<i>Zhao Qu, Shanshan Zhang and Kun Ding</i>	
Who Files Provisional Applications in the United States?	838
<i>Chi-Tung Chen and Dar-Zen Chen</i>	
A Preliminary Study of Technological Evolution: From the Perspective of the USPC Reclassification	847
<i>Hui-Yun Sung, Chun-Chieh Wang and Mu-Hsuan Huang</i>	
Cognitive Distances in Prior Art Search by the Triadic Patent Offices: Empirical Evidence from International Search Reports	859
<i>Tetsuo Wada</i>	
A Collective Reasoning on the Automotive Industry: A Patent Co-citation Analysis	865
<i>Manuel Castriotta and Maria Chiara Di Guardo</i>	
Statistical Study of Patents Filed in Global Nano Photonic Technology	871
<i>Zhang Huijing, Zhong Yongheng and Jiang Hong</i>	
A Sao-Based Approach for Technologies Evolution Analysis Using Patent Information: Case Study on Graphene Sensor	873
<i>Zhengyin Hu and Shu Fang</i>	
Prediction of Potential Market Value Using Patent Citation Index	875
<i>HeeChel Kim, Hong-Woo Chun and Byoung-Youl Coh</i>	
Knowledge Flows and Delays in the Pharmaceutical Innovation System	877
<i>Mari Jibu, Yoshiyuki Osabe, and Katy Börner</i>	

THEORY AND METHODS & TECHNIQUES	PAGE
Can Numbers of Publications on a Specific Topic Observe the Research Trend of This Topic: A Case Study of the Biomarker HER-2?	879
<i>Yuxian Liu Michael Hopkins and Yishan Wu</i>	
Founding Concepts and Foundational Work: Establishing the Framework for the Use of Acknowledgments as Indicators	890
<i>Nadine Desrochers, Adèle Paul-Hus and Jen Pecoskie</i>	
Analysis On The Age Distribution Of Scientific Elites' Productivity: A Study On Academicians Of The Chinese Academy Of Science	895
<i>Liu Jun-wan, Zheng Xiao-min, Feng Xiu-zhen and Wang Fei-fei</i>	
An Experimental Study On The Dynamic Evolution Of Core Documents	897
<i>Lin Zhang, Wolfgang Glänzel and Fred Y. Ye</i>	
How Related is Author Topical Similarity to Other Author Relatedness Measures?	899
<i>Kun Lu, Yuehua Zhao, Isola Ajiferuke and Dietmar Wolfram</i>	

Publication Rates in 192 Research Fields of the Hard Sciences	909
<i>Ciriaco Andrea D'Angelo and Giovanni Abramo</i>	
A Technology Foresight Model: Used for Foreseeing Impelling Technology in Life Science	920
<i>Yunwei Chen, Yong Deng, Fang Chen, Chenjun Ding, Ying Zheng and Shu Fang</i>	
Lung Cancer Researchers, 2008-2013: Their Sex and Ethnicity	932
<i>Grant Lewison, Philip Roe and Richard Webber</i>	
A Model for Publication and Citation Statistics of Individual Authors	942
<i>Wolfgang Glänzel, Sarah Heeffner and Bart Thijs</i>	
A Delineating Procedure to Retrieve Relevant Research Areas on Nanocellulose	953
<i>Douglas H. Milanez and Ed C. M. Noyons</i>	
Sapientia: the Ontology of Multi-dimensional Research Assessment	965
<i>Cinzia Daraio, Maurizio Lenzerini, Claudio Leporelli, Henk F. Moed, Paolo Naggar, Andrea Bonaccorsi and Alessandro Bartolucci</i>	
The Research Purpose, Methods and Results of the "Annual Report for International Citations of China's Academic Journals"	978
<i>Junhong Wu, Hong Xiao, Shuhong Sheng, Yan Zhang, Xiukun Sun and Yichuan Zhang</i>	
Is the Year of First Publication a Good Proxy of Scholars' Academic Age?	988
<i>Rodrigo Costas, Tina Nane and Vincent Larivière</i>	
Corpus Specific Stop Words to Improve the Textual Analysis in Scientometrics	999
<i>Vicenç Parisi Baradad and Alexis-Michel Mugabushaka</i>	
Epistemic Diversity as Distribution of Paper Dissimilarities	1006
<i>Jochen Gläser, Michael Heinz and Frank Havemann</i>	
Using Bibliometrics-aided Retrieval to Delineate the Field of Cardiovascular Research	1018
<i>Diane Gal, Karin Sipido and Wolfgang Glänzel</i>	
Locating an Astronomy and Astrophysics Publication Set in a Map of the Full Scopus Database	1024
<i>Kevin W. Boyack</i>	
Scientific Workflows for Bibliometrics	1029
<i>Arzu Tugce Guler, Cathelijn J. F. Waaijer and Magnus Palmblad</i>	
Expertise Overlap between an Expert Panel and Research Groups in Global Journal Maps	1035
<i>A.I.M. Jakaria Rahman, Raf Guns, Ronald Rousseau and Tim C.E. Engels</i>	
Contextualization of Topics - Browsing through Terms, Authors, Journals and Cluster Allocations	1042
<i>Rob Koopman, Shenghui Wang and Andrea Scharnhorst</i>	
A Link-based Memetic Algorithm for Reconstructing Overlapping Topics from Networks of Papers and their Cited Sources	1054
<i>Frank Havemann, Jochen Gläser and Michael Heinz</i>	
Re-citation Analysis: A Promising Method for Improving Citation Analysis for Research Evaluation, Knowledge Network Analysis, Knowledge Representation and Information Retrieval	1061
<i>Dangzhi Zhao and Andreas Strotmann</i>	
Topic Affinity Analysis for an Astronomy and Astrophysics Data Set	1066
<i>Theresa Velden, Shiyan Yan and Carl Lagoze</i>	

Time & Citation Networks	1073
<i>James R. Clough and Tim S. Evans</i>	
Coming to Terms: A Discourse Epistemetrics Study of Article Abstracts from the Web of Science	1079
<i>Bradford Demarest, Vincent Larivière and Cassidy R. Sugimoto</i>	
Using Hybrid Methods and ‘Core Documents’ for the Representation of Clusters and Topics: The Astronomy Dataset	1085
<i>Wolfgang Glänzel and Bart Thijs</i>	
Mining Scientific Papers for Bibliometrics: A (Very) Brief Survey of Methods and Tools	1091
<i>Iana Atanassova, Marc Bertin and Philipp Mayr</i>	
A Multi-Agent Model of Individual Cognitive Structures and Collaboration In Sciences	1093
<i>Bulent Ozel</i>	
Hypothesis Generation for Joint Attention Analysis on Autism	1095
<i>Jian Xu, Ying Ding, Chaomei Chen and Erjia Yan</i>	
“What Came First – Wellbeing Or Sustainability?” A Systematic Analysis of The Multi-Dimensional Literature Using Advanced Topic Modelling Methods	1097
<i>Mubashir Qasim and Les Oxley</i>	
Multi-Label Propagation for Overlapping Community Detection Based on Connecting Degree	1099
<i>Xiaolan Wu and Chengzhi Zhang</i>	
Reproducibility, Consensus and Reliability In Bibliometrics	1101
<i>Raul I. Mendez-Vasquez and Eduard Suñen-Pinyol</i>	
Semantometrics: Fulltext-Based Measures for Analysing Research Collaboration	1103
<i>Drahomira Herrmannova and Petr Knoth</i>	
Uncovering the Mechanisms of Co-Authorship Network Evolution by Multirelations-Based Link Prediction	1105
<i>Jinzhu Zhang, Chengzhi Zhang and Bikun Chen</i>	

JOURNALS, DATABASES AND ELECTRONIC PUBLICATIONS / DATA ACCURACY AND DISAMBIGUATION / MAPPING AND VISUALIZATION	PAGE
Citing e-prints on arXiv A study of cited references in WoS-indexed journals from 1991-2013	1107
<i>Valeria Aman</i>	
Evolutionary Analysis of Collaboration Networks in <i>Scientometrics</i>	1121
<i>Yuehua Zhao and Rongying Zhao</i>	
Open Access Publishing and Citation Impact - An International Study	1130
<i>Theo van Leeuwen, Clifford Tatum and Paul Wouters</i>	
Measuring the Competitive Pressure of Academic Journals and the Competitive Intensity within Subjects	1142
<i>Ma Zheng, Pan Yuntao, Wu Yishan, Yu Zhenglu and Su Cheng</i>	
SciELO Citation Index and Web of Science: Distinctions in the Visibility of Regional Science	1152
<i>Diana Lucio-Arias, Gabriel Velez-Cuartas and Loet Leydesdorff</i>	

Book Bibliometrics – A New Perspective and Challenge in Indicator Building Based on the Book Citation Index	1161
<i>Pei-Shan Chi, Wouter Jeuris, Bart Thijs and Wolfgang Glänzel</i>	
When is an Article Actually Published? An Analysis of Online Availability, Publication, and Indexation Dates	1170
<i>Stefanie Haustein, Timothy D. Bowman and Rodrigo Costas</i>	
Analysis of the Obsolescence of Citations and Access in Electronic Journals at University Libraries	1180
<i>Chizuko Takei, Fuyuki Yoshikane and Hiroshi Itsumura</i>	
Dynamics Between National Assessment Policy and Domestic Academic Journals	1191
<i>Eleonora Dagienė and Ulf Sandström</i>	
Correlation between Impact Factor and Public Availability of Published Research Data in Information Science & Library Science Journals	1194
<i>Rafael Aleixandre-Benavent, Luz Moreno-Solano, Antonia Ferrer Sapena and Enrique Alfonso Sánchez Pérez</i>	
Use of CrossRef and OAI-PMH to Enrich Bibliographical Databases	1196
<i>Mehmet Ali Abdulhayoglu and Bart Thijs</i>	
Does Scopus Really Put Journal Selection Criteria into Practice?	1198
<i>Zehra Taşkın, Güleda Doğan, Sümeyye Akça, İpek Şencan and Müge Akbulut</i>	
On the Correction of “Old” Omitted Citations by Bibliometric Databases	1200
<i>Fiorenzo Franceschini, Domenico Maisano and Luca Mastrogiacomo</i>	
Can We Track the Geography of Surnames Based on Bibliographic Data?	1208
<i>Nicolas Robinson-Garcia, Ed Noyons and Rodrigo Costas</i>	
An 80/20 Data Quality Law for Professional Scientometrics?	1218
<i>Andreas Strotmann and Dangzhi Zhao</i>	
Some Features of the Citation Counts from Journals Indexed in Web of Science to Publications from Russian Translation Journals	1220
<i>Maria Aksenteva</i>	
Semantics, A Key Concept in Interoperability of Research Information -The Flanders Research Funding Semantics Case	1222
<i>Sadia Vancauwenbergh</i>	
The Information Retrieval Process of the Scientific Production at Departmental-Level of Universities: Exploration of New Approach	1224
<i>César David Loaiza Quintana and Víctor Andrés Bucheli Guerrero</i>	
Efficiency, Effectiveness and Impact of Research and Innovation: A Framework for the Analysis	1226
<i>Cinzia Daraio</i>	
Integrating Microdata on Higher Education Institutions (HEIS) with Bibliometric and Contextual Variables: A Data Quality Approach	1228
<i>Cinzia Daraio, Angelo Gentili and Monica Scannapieco</i>	
Is The Humboldtian University Model An Engine Of Local Development? New Empirical Evidence From The ETER Database	1230
<i>Teresa Ciorciaro, Libero Cornacchione, Cinzia Daraio and Giulia Dionisio</i>	

Connecting Big Scholarly Data With Science Of Science Policy: An Ontology-Based-Data-Management (OBDM) Approach	1232
<i>Cinzia Daraio¹, Maurizio Lenzerini, Claudio Leporelli, Henk F. Moed, Paolo Naggar, Andrea Bonaccorsi and Alessandro Bartolucci</i>	
Incomplete Data and Technological Progress in Energy Storage Technologies	1234
<i>Sertaç Oruç, Scott W. Cunningham, Christopher Davis and Bert van Dorp</i>	
Bibliometric Characteristics of a “Paradigm Shift”: The 2012 Nobel Prize in Medicine	1244
<i>Andreas Strotmann and Dangzhi Zhao</i>	
Bibliometric Mapping: Eight Decades of Analytical Chemistry, With Special Focus on the Use of Mass Spectrometry	1250
<i>Cathelijan J. F. Waaijer and Magnus Palmblad</i>	
Introduction of “Kriging” to Scientometrics for Representing Quality Indicators in Maps of Science	1252
<i>Masashi Shirabe</i>	
The Technology Roots Spectrum: A New Visualization Tool for Identifying the Roots of a Technology	1255
<i>Eduardo Perez-Molina</i>	
Modelling of Scientific Collaboration Based on Graphical Analysis	1257
<i>Veslava Osinska, Grzegorz Osinski and Wojciech Tomaszewski</i>	
Monitoring of Technological Development - Detection of Events in Technology Landscapes through Scientometric Network Analysis	1259
<i>Geraldine Joanny, Adam Agocs, Sotiri Fragkiskos, Nikolaos Kasfikis, Jean-Marie Le Goff and Olivier Eulaerts</i>	
Analysis of R&D Trend for the Treatment of Autoimmune Diseases by Scientometric Method	1261
<i>Eunsoo Sohn, Oh-Jin Kwon, Eun-Hwa Sohn and Kyung-Ran Noh</i>	
Analysis of Convergence Trends in Secondary Batteries	1263
<i>Young-Duk Koo and Dae-Hyun Jeong</i>	
Can Scholarly Literature and Patents be Represented in a Hierarchy of Topics Structured to Contain 20 Topics per Level? Balancing Technical Feasibility with Human Usability	1265
<i>Michael Edwards, Mahadev Dovre Wudali, James Callahan, Paul Worner, Jeffrey Maudal, Patricia, Brennan, Julia Laurin and Joshua Schnell</i>	
A Sciento-Text Framework for Fine-Grained Characterization of the Leading World Institutions in Computer Science Research	1267
<i>Ashraf Uddin, Sumit Kumar Banshal, Khushboo Singhal and Vivek Kumar Singh</i>	
Influence of Human Behaviour and the Principle of Least Effort on Library and Information Science Research	1269
<i>Yu-Wei Chang</i>	
Document Type Assignment Accuracy in Citation Index Data Sources	1271
<i>Paul Donner</i>	
Measuring the Impact of Arabic Scientific Publication: Challenges and Proposed Solution	1273
<i>Raad Alturki</i>	

On the Correction of “Old” Omitted Citations by Bibliometric Databases

Fiorenzo Franceschini¹, Domenico Maisano² and Luca Mastrogiacomo³

¹*fiorenzo.franceschini@polito.it*, ²*domenico.maisano@polito.it*, ³*luca.mastrogiacomo@polito.it*
Politecnico di Torino, DIGEP (Department of Management and Production Engineering),
Corso Duca degli Abruzzi 24, 10129, Torino (Italy)

Abstract

Omitted citations – i.e., missing links between a cited paper and the corresponding citing papers – are the main consequence of several bibliometric-database errors. To reduce these errors, databases may undertake two actions: (i) improving the control of the (new) papers to be indexed, i.e., limiting the introduction of “new” dirty data, and (ii) detecting and correcting errors in the papers already indexed by the database, i.e., cleaning “old” dirty data. The latter action is probably more complicated, as it requires the application of suitable error-detection procedures to a huge amount of data. Based on an extensive sample of scientific papers in the Engineering-Manufacturing field, this study focuses on old dirty data in the Scopus and WoS databases. To this purpose, a recent automated algorithm for estimating the omitted-citation rate of databases is applied to the same sample of papers, but in three different-time sessions. A database’s ability to clean the old dirty data is evaluated considering the variations in the omitted-citation rate from session to session. The major outcomes of this study are that: (i) both databases slowly correct old omitted citations, and (ii) a small portion of initially corrected citations can surprisingly come off from databases over time.

Conference Topic

Data Accuracy and disambiguation

Introduction

An important branch of the bibliometric literature examines errors in bibliometric databases. Several studies show that the major consequence of database errors is represented by omitted citations, i.e., citations that should be ascribed to a certain (cited) paper but, for some reason, are lost (Moed, 2005; Buchanan, 2006; Jacsó, 2006; Li et al., 2010; Olensky, 2013).

Franceschini et al. (2013) proposed an automated algorithm for estimating the omitted-citation rate of bibliometric databases. This algorithm requires the combined use of two or more bibliometric databases and is based upon the hypothesis that the mismatch between the citations occurring in one database and another one is evidence of possible errors/omissions.

In a further study by Franceschini et al. (2014), this algorithm was applied to a relatively large set of publications, showing that, depending on the bibliometric database in use (Scopus or WoS), omitted citations are not distributed uniformly among publishers; e.g., regarding the publications in the Engineering-Manufacturing field, citations from papers published by Wiley-Blackwell are more likely to be omitted by Scopus, while those from papers published by ASME (American Society of Mechanical Engineers) are more likely to be omitted by WoS. A reason behind this result is that some editorial styles imposed by certain publishers can probably hamper the correct identification of the cited papers by some databases.

The presence of database errors, as well as journal coverage or author disambiguation, is probably one of the major concerns of database administrators. In the authors’ opinion, database administrators may undertake two actions for reducing database errors:

1. Limiting the introduction of “new” dirty data in a database, i.e., errors concerning new papers to be indexed;
2. Cleaning “old” dirty data, i.e., errors concerning papers/journals already indexed by a database.

The recent effort by reviewers, publishers and database administrators in checking the cited article lists of new papers probably contributes to reducing “new” dirty data. This hypothesis is corroborated by a recent study by Franceschini et al. (2015), which shows that the databases’ propensity to omit newer citations is generally lower than that to omit older citations.

Cleaning up old dirty data is certainly much more complicated because it requires the systematic application of suitable error-detection procedures to a huge amount of data. However, this effort would be essential for improving the quality of a database significantly.

This paper focuses on the ability of the major multidisciplinary bibliometric databases, i.e., Scopus and WoS, to clean up old dirty data. For this evaluation, we use a new procedure, derived from the automated algorithm by Franceschini et al. (2013). This procedure consists in (i) repeating the omitted-citation-rate analysis on the same sample of (cited and citing) articles, but in different-time sessions, and (ii) observing any variation in the results. A database’s ability to clean old dirty data will be evaluated considering the variation in the omitted-citation rate from one session to another one.

The remainder of this paper is organized into four sections. The section “Automated algorithm for examining the omitted citations” briefly recalls the algorithm by Franceschini et al. (2013). The section “Methodology” describes the methodology used in our study, focusing on data collection and analysis. The section “Results” illustrates the results of the analysis, investigating similarities and differences between the two databases examined. Finally, the section “Conclusions” summarizes the original contributions of this paper, highlighting the major results, limitations and suggestions for future research.

Automated algorithm for analysing the omitted citations

Before recalling the algorithm, we present an introductory example to illustrate how it works. Let us consider a fictitious paper of interest, indexed by Scopus and WoS. The number of citations received by this paper is four in Scopus and six in WoS (see Table 1).

Table 1. Citation data relating to a fictitious article, according to Scopus and WoS. The union of the citations recorded by the two databases (see the first column) is a total of eight citations. Among the citations, only five come from sources officially covered by both databases (highlighted in grey).

Citation No.	Scopus	WoS
1	✓	
2		✓
3	Omitted	✓
4	✓	✓
5	✓	✓
6	Omitted	✓
7		✓
8	✓	Omitted
Total	4	6

The union of the citations recorded by the two databases is a total of eight citations. Among the citations, only five come from sources (i.e., journals or conference proceedings) officially covered by both databases (highlighted in grey in Table 1). Focusing on these five “theoretically overlapping” (TO) citations, two are omitted by Scopus (but not by WoS) and one is omitted by WoS (but not by Scopus). Therefore, from the perspective of the paper of interest, a rough estimate of the omitted-citation rate is $2/5 \approx 40\%$ in Scopus and $1/5 \approx 10\%$ in WoS. The same reasoning can be extended to multiple papers of interest and more than two bibliometric databases.

The automated algorithm, which is based on the combined use of two bibliometric databases (Scopus and WoS in this case), can be summarised in three steps:

1. Identify a set of (P) papers of interest, indexed by both the databases.
2. For each (i -th) paper of the set, identify the TO citations, defined as the portion of documents issued by journals officially covered by Scopus and WoS. The number of TO citations concerning the i -th paper of interest will be denoted as γ_i .
3. For each (i -th) paper of the set and for each database, determine the number (ω_i) of TO citations that do not occur in it and classify them as omitted citations. The omitted-citation rate (p) relating to the P papers of interest, according to a database, can be estimated as:

$$\hat{p} = \sum_{i=1}^P \omega_i / \sum_{i=1}^P \gamma_i. \quad (1)$$

We emphasize that p is estimated on the basis of (i) a set of papers of interests and (ii) a portion of the total citations that they obtained (i.e., that ones related to citing articles purportedly covered by both the databases). For a more detailed description of the algorithm, we refer the reader to Franceschini et al. (2013).

The ability of bibliometric databases to clean old dirty data will be evaluated by applying this algorithm to the same sample of TO citations, in three different-time sessions.

Methodology

The study is based on the analysis of the citations obtained from a relatively large sample of papers of interest. The papers were issued by 33 scientific journals (i) included in the ISI Subject Category of Engineering-Manufacturing (by WoS) and (ii) covered by Scopus; Table 2 reports the list of these journals. For each journal, we considered the papers published in the time-window from 2006 to 2012 and the citations that they obtained from papers issued in the same period.

Data collection was repeated in three different-time sessions, spaced about seven months apart: i.e., session I on August 2013, session II on March 2014 and session III on September 2014. We remark that the duration of each data-collection session (i.e., a few days) is negligible with respect to the time period between two consecutive sessions.

To enable comparisons between data collected in different sessions, we adopted two measures:

1. Among the papers of interest (or cited papers) – i.e., those issued by the 33 Engineering-Manufacturing journals – we selected those indexed in each of the three sessions, by both the (Scopus and WoS) databases; in formal terms:

$$A = A^{(I)} \cap A^{(II)} \cap A^{(III)}, \quad (2)$$

A being the set of cited papers selected for our analysis and $A^{(I)}$, $A^{(II)}$ and $A^{(III)}$ the sets of papers indexed by both the databases, at the moment of session I, II and III respectively.

Also, we excluded articles without DOI code or whose DOI code is not indexed by both databases, as they would be difficult to disambiguate.

2. Among the citations, we selected the so-called TO citations, i.e., those obtained from journals purportedly covered by both databases and issued in the 2006-to-2012 time-window. To avoid any misunderstanding, we excluded citations from journals covered in the 2006-to-2012 time-window, but later banned from the database¹. The official lists of documents covered by the databases in use – which are essential for determining the TO

¹ A possible misunderstanding arises from the fact that, in some cases (mostly on Scopus), the expulsion of a journal from a database entails the entire removal of previously indexed papers, while in other cases (mostly on WoS), previously indexed papers are not necessarily removed.

citations – were retrieved from the databases’ websites (Scopus Elsevier, 2015; Thomson Reuters, 2015).

Table 2. List of the Engineering-Manufacturing journals examined. For each journal, it is reported its title and ISSN code. Journals are sorted alphabetically according to their title

Journal title	ISSN
AI EDAM - Artificial Intelligence for Engineering Design Analysis and Manufacturing	0890-0604
Assembly Automation	0144-5154
CIRP Annals - Manufacturing Technology	0007-8506
Composites Part A - Applied Science and Manufacturing	1359-835X
Concurrent Engineering - Research and Applications	1063-293X
Design Studies	0142-694X
Flexible Services and Manufacturing Journal	1936-6582
Human Factors and Ergonomics in Manufacturing & Service Industries	1090-8471
IEEE Transaction on Components Packaging and Manufacturing Technology	2156-3950
IEEE Transactions on Semiconductor Manufacturing	0894-6507
IEEE-ASME Transactions on Mechatronics	1083-4435
International Journal of Advanced Manufacturing Technology	0268-3768
International Journal of Computer Integrated Manufacturing	0951-192X
International Journal of Crashworthiness	1358-8265
International Journal of Machine Tools & Manufacture	0890-6955
International Journal of Production Economics	0925-5273
Journal of Advances Mechanical Design Systems and Manufacturing	1881-3054
Journal of Computing and Information Science in Engineering - Transactions of the ASME	1530-9827
Journal of Intelligent Manufacturing	0956-5515
Journal of Manufacturing Science and Engineering - Transactions of the ASME	1087-1357
Journal of Manufacturing Systems	0278-6125
Journal of Materials Processing Technology	0924-0136
Journal of Scheduling	1094-6136
Machining Science and Technology	1091-0344
Materials and Manufacturing Processes	1042-6914
Proceedings of the Institution of Mechanical Engineers Part B - Journal of Engineering Manufacture	0954-4054
Packaging Technology and Science	0894-3214
Precision Engineering - Journal of the International Societies for Precision Engineering and Nanotechnology	0141-6359
Production and Operations Management	1059-1478
Production Planning & Control	0953-7287
Research in Engineering Design	0934-9839
Robotics and Computer-Integrated Manufacturing	0736-5845
Soldering & Surface Mount Technology	0954-0911

The sample of TO citations used in the analysis is the union of the TO citations (that meet the above requirements), collected in each of the three sessions. In formal terms, this sample of TO citations is:

$$B = B^{(I)} \cup B^{(II)} \cup B^{(III)}, \quad (3)$$

$B^{(I)}$, $B^{(II)}$ and $B^{(III)}$ being the TO citations collected during session I, II and III respectively.

This sample of TO citations will be used for estimating the omitted-citations rate of a certain database, in a certain session; the relationship in Eq. 1 can be used, being:

\hat{p} the estimate of the omitted-citation rate related to a certain session and a specific database;

P the number of (cited) articles of interest;

γ_i the number of TO citations relating to the i -th of the P articles of interest;

ω_i the portion of the TO citations, collected in a certain session, which are omitted by a specific database.

Being \hat{p} just an estimate of p – albeit the best possible – a relevant symmetrical $(1 - \alpha)$ confidence interval (*CI*) can be constructed as²:

$$\hat{p} \pm z_{1-\alpha/2} \sqrt{\frac{\hat{p} \cdot (1 - \hat{p})}{\sum_{i=1}^P \gamma_i}}, \quad (4)$$

with:

α , the type-I error;

$z_{1-\alpha/2}$ the unit normal deviate corresponding to $1 - \alpha/2$.

In this case, we consider a symmetrical 95% *CI*, therefore $\alpha = 5\%$ and $z_{97.5\%} \approx 2$.

By adopting this procedure, we will obtain six different estimates of the omitted-citation rate, i.e., one for each of the three sessions and each of the two databases in use. The comparison of these estimates will tell us whether the databases examined are able to correct old omitted citations.

Results

The total number of papers of interest, i.e., those issued by the Engineering-Manufacturing journals examined, is $P = 23,806$. The corresponding TO citations are $\sum \gamma_i = 97,698$. Table 3 contains the \hat{p} values and the relevant 95% *CI*s, relating to the three sessions and the two databases examined.

Table 3. Main results of the (repeated) analysis of the omitted-citation rate of databases. Citing and cited articles were issued from 2006 to 2012. Statistics concern each of the three sessions (i.e., session I, II and III) for Scopus and WoS respectively.

Session	$\sum_{i=1}^P \gamma_i$	(a) Scopus				(b) Wos			
		$\sum_{i=1}^P \omega_i$	\hat{p}	95% <i>CI</i>		$\sum_{i=1}^P \omega_i$	\hat{p}	95% <i>CI</i>	
I (August 2013)	97,698	5,183	5.3%	5.2%	5.4%	7,370	7.5%	7.4%	7.7%
II (March 2014)	97,698	4,607	4.7%	4.6%	4.8%	6,376	6.5%	6.4%	6.7%
III (October 2014)	97,698	4,473	4.6%	4.4%	4.7%	6,404	6.6%	6.4%	6.7%

$P = 97,698$ is the total number of (cited) articles, published by 33 Engineering-Manufacturing journals;

$\sum \gamma_i$ is the total number of TO citations (which is independent on the session);

$\sum \omega_i$ is the total number of omitted citations relating to each session and each database;

\hat{p} is the estimate of the omitted-citation rate relating to each session and each database;

The 95% *CI* around \hat{p} is obtained applying the approximated relationship in Eq. 4.

² The *CI* construction in Eq. 4 is grounded on the following considerations:

- For a generic sample consisting of $n = \sum \gamma_i$ TO citations, the number of omitted citations will be a binomially distributed variable with mean value $n \cdot p$ and variance $n \cdot p \cdot (1 - p)$;
- The aforesaid binomial distribution can be approximated by a normal distribution with the same mean value and variance. This approximation is acceptable in the case $n \cdot p \geq 5$ (Ross, 2009), which is generally satisfied when considering relatively large sets of TO citations.
- Based on the previous approximation, the percentage of omitted citations for a sample of n TO citations will be a normally distributed variable with mean value p and variance $p \cdot (1 - p) / n$. Since p is not known, it can be replaced by its best estimate \hat{p} .

In conclusion, Eq. 4 defines a symmetric *CI* around \hat{p} , which – with a probability $(1 - \alpha)$ – will include the “true” p value.

The \hat{p} values of both databases tend to decrease over time, denoting that dirty data have been partially cleaned. Interestingly, the major reduction in the \hat{p} values is between the session I and II for both databases; on the other hand, variations between session II and III are not significant, since the 95% CIs are partially overlapped (see Figure 1(a)); as regards WoS, we can even notice an imperceptible increase in the \hat{p} value between session II and III.

The overall reduction in the number of omitted TO citations ($\sum \omega_i$) for WoS is greater than that for Scopus (i.e., $7,370 - 6,404 = 966$ against $5,183 - 4,473 = 710$); however, consistently with what observed in other studies (Franceschini et al., 2014; 2015), we note that the omitted-citation rates in Scopus are generally lower than those in WoS. Figure 1(b) shows that the overall percent variations in the \hat{p} values between session I and III are very similar (i.e., -13.7% and -13.1%, for Scopus and WoS respectively).

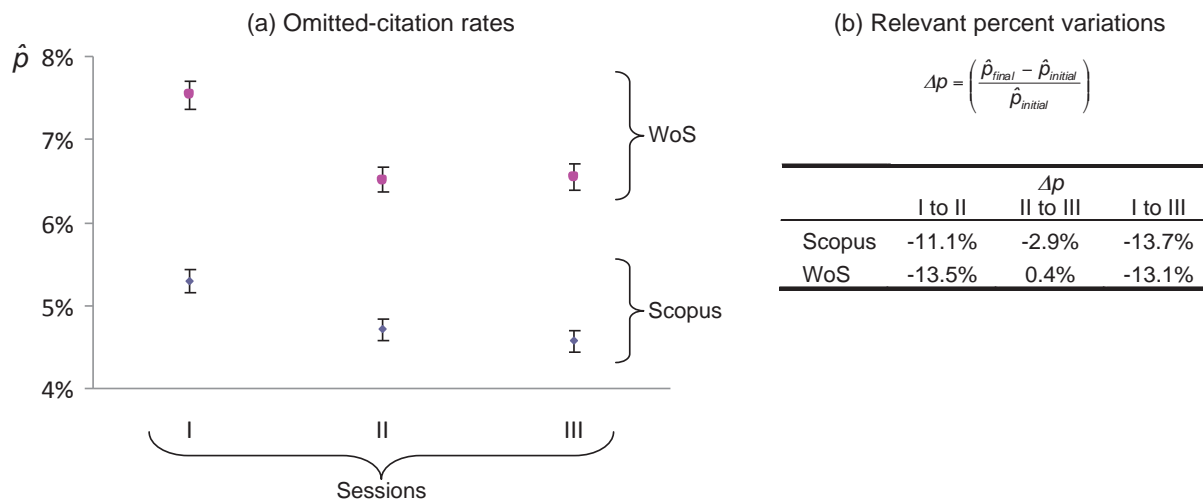


Figure 1. (a) Graphical representation of the omitted-citation rate in the three sessions, for Scopus and WoS, and (b) relevant percent variations.

Having verified that both databases tend to slowly correct old omitted citations, we now investigate the possible differences in the indexing of individual TO citations, from one session to another one. Table 4 summarizes the eight possible events concerning the correct/missing indexing of individual TO citations. Since there are two possible indexing states (i.e., correct or missing indexing) for each of the three sessions, the total number of possible events is $2^3 = 8$; the file containing the complete list of individual TO citations, with the relevant cited papers, and their session-by-session indexing by the databases, is available under request to authors.

Not surprisingly, the most frequent events are those with no variation (i.e., the type 1 and 2 events in Table 4), in which the TO citations are indexed correctly (“√”) or incorrectly (“×”) in all the three sessions; the portion of TO citations with no variation is 98.7% for Scopus and 98.5% for WoS). The type 3 and 4 events represent corrections in the TO-citation indexing, in session II and III respectively. The total number of corrections in WoS is basically larger to that in Scopus, probably due to the larger level of “initial dirt” in the former database, compared to that one in the latter. Moreover, we note that almost all of the corrections by WoS are concentrated in session II (i.e., 1193 out of 1215).

Despite these differences, the percentage of TO citations corrected by Scopus and WoS are pretty close to each other (i.e., roughly 1% and 1.2% respectively). This similarity is even more interesting if we consider the fact that, among the set of corrected TO citations, a relatively small subset is shared between the two databases (i.e., 392 citations out of $(997 + 1,215 - 392) = 1,820$, corresponding to about 21.5% of the set of corrected TO citations).

Table 4. Overall statistics concerning the indexing of the individual TO citations, in each session. Symbols “✓” and “✗” respectively identify the TO citations correctly indexed or omitted in a certain session.

Type of event	Session			(a) Scopus				(b) WoS				
	I	II	III	Single event		Aggregated events		Single event		Aggregated events		
				TO citations	Percent	TO citations	Percent	TO citations	Percent	TO citations	Percent	
No variation	1	✓	✓	✓	92,296	94.5%	96,411	98.7%	90,195	92.3%	96,214	98.5%
	2	✗	✗	✗	4,115	4.2%			6,019	6.2%		
Correction	3	✗	✓	✓	765	0.8%	997	1.0%	1,193	1.2%	1,215	1.2%
	4	✗	✗	✓	232	0.2%			22	0.0%		
Anomalous variation	5	✓	✗	✗	102	0.1%	290	0.3%	164	0.2%	269	0.3%
	6	✓	✓	✗	112	0.1%			77	0.1%		
	7	✗	✓	✗	0	0.0%			0	0.0%		
	8	✓	✗	✓	76	0.1%			28	0.0%		
Total				97,698	100%	97,698	100%	97,698	100%	97,698	100%	

The type 5 to 8 events are characterized by anomalous variations, in which some TO citations, which are correctly indexed in a certain session, are omitted in one (or more) subsequent sessions. It is surprising how citations, which were initially indexed correctly, can come off from a database over time; in other words, these events represent a form of generation of dirty data, which is independent of the introduction of new data in the database. Fortunately, the incidence of these abnormalities is rather low (coincidentally, about 0.3% for both Scopus and for WoS); in the future, we may conduct a thorough analysis of these anomalies, based on their manual examination.

Conclusions

The analysis presented in this paper shows that the two bibliometric database examined tend to gradually reduce the number of old omitted citations, although this reduction is relatively slow for both. It would be interesting to see to what extent these cleanings were due to error-correction campaigns structured by database administrators, or simply due to impromptu database-inaccuracy reports by authors and/or database users (even checking and cleaning up bibliometric data in personal research profiles, such as ResearcherID, Scopus Author ID, ORCID, etc.).

Results of this study show other interesting similarities/coincidences between the two databases examined:

1. Comparing the results related to session I and III (spaced about fourteen months apart), we noticed a 13-to-14% reduction in the p values for both Scopus and WoS.
2. For both databases, the greatest reduction in the omitted-citations rate was registered in session II and not in session III. This could be just a coincidence or it could denote a sort of “seasonality” of the two databases in cleaning up old dirty data.
3. The portion of TO citations whose indexing varies in the three sessions is roughly the same for both databases, i.e., roughly 1 to 1.5%. Apart from the previously omitted TO citations that have been justly corrected, they include a small portion of abnormal variations, i.e., TO citations correctly indexed in some session and subsequently omitted. Coincidentally, the percentage of abnormal variations is 0.3% for both databases.

The proposed analysis has several limitations. Even though the set of TO citations includes almost one-hundred thousand citations, the relevant cited papers are all confined within the Engineering-Manufacturing field. Also, the analysis was repeated in three sessions over a

total period of about 14 months; therefore, it reflects a database's ability to correct errors in short/middle-term period, but not in the long-term period.

In the future, we plan to extend the study to a longer time-scale (e.g., 2 or 3 years) and/or to scientific articles in other disciplines. Furthermore, the study will be expanded for investigating possible links between the omitted citations' propensity to be corrected and the publishers of the relevant citing papers.

References

- Buchanan, R.A. (2006). Accuracy of Cited References: The Role of Citation Databases. *College & Research Libraries*, 67(4), 292-303.
- 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. (2015). Influence of omitted citations on the bibliometric statistics of the major Manufacturing journals. To appear in *Scientometrics*. A draft version is available at http://staff.polito.it/fiorenzo.franceschini/Pubblicazioni/Revised_IJPE-D-13-01272.pdf.
- Jacsó, P. (2006). Deflated, inflated and phantom citation counts. *Online Information Review*, 30(3), 297-309.
- Li, J., Burnham, J.F., Lemley, T., & Britton, R.M. (2010). Citation analysis: comparison of Web of Science, Scopus, Scifinder, and Google Scholar. *Journal of Electronic Resources in Medical Libraries* 7(3), 196-217.
- Moed, H.F. (2006). *Citation analysis in research evaluation* (Vol. 9). Springer.
- Olensky, M. (2013). Accuracy Assessment for Bibliographic Data. *Proceedings of the 13th International Conference of the International Society for Scientometrics and Informetrics (ISSI)*, vol. 2, pp. 1850-1851, Vienna, Austria.
- Ross, S.M. (2009). *Introduction to probability and statistics for engineers and scientists*. Academic Press.
- Schenker, N., & Gentleman, J.F. (2001). On judging the significance of differences by examining the overlap between confidence intervals. *The American Statistician*, 55(3), 182-186.
- Scopus Elsevier (2015). *Scopus Content Coverage*. Available at <http://www.scopus.com> [retrieved on August 2013, March 2014 and October 2014].
- Thomson Reuters (2015). *Master Journal List*, <http://ip-science.thomsonreuters.com/mjl/> [retrieved on August 2013, March 2014 and October 2014].