21st international conference on science and technology indicators

STI Conference 2016 · València

Peripheries, frontiers and beyond

14 ·16 September 2016

Universitat Politècnica de València

BOOK OF PROCEEDINGS

www.sti2016.org

Congress UPV
Proceedings of the 21ST International Conference on Science and Technology Indicators

The contents of this publication have been evaluated by the Scientific Committee which it relates and the procedure set out http://ocs.editorial.upv.es/index.php/STI2016/STI2016

Scientific Editors
Ismael Ràfols
Jordi Molas-Gallart
Elena Castro-Martínez
Richard Woolley

ISBN: 978-84-9048-519-4 (print version)
Print on-demand

DOI: http://dx.doi.org/10.4995/STI2016.2016.4543

TABLE OF CONTENTS

Page

ORAL PRESENTATIONS

CHAPTER 1: Data Infrastructure and Data Quality

CHAPTER 2: Smart Use of Indicators for Innovation Policy

CHAPTER 3: Measuring Innovation

CHAPTER 4: Societal Impact

CHAPTER 5: Mission Oriented Research

CHAPTER 6: Rethinking Evaluation

CHAPTER 7: Programme and Project Assessment

CHAPTER 8: Collaborations, Mobility and Internationalisation

CHAPTER 9: Careers, Labour Market and Individual Performance

CHAPTER 10: Gender

CHAPTER 11: Social Sciences and Humanities

CHAPTER 12: Society, Participation and Culture

CHAPTER 13: Indicators in Regional and National Contexts

CHAPTER 14: Networks and Text Analysis

CHAPTER 15: Altmetrics

CHAPTER 16: Citation Impact

POSTER SESSION 1: Indicators, Assessment, Funding and Innovation

POSTER SESSION 2: Data Characterisation, Classification, Visualisation and Indicator Design

INTRODUCTION

This is the book of proceedings of the 21st Science and Technology Indicators Conference that took place in València (Spain) from 14th to 16th of September 2016.

The conference theme for this year, ‘Peripheries, frontiers and beyond’ aimed to study the development and use of Science, Technology and Innovation indicators in spaces that have not been the focus of current indicator development, for example, in the Global South, or the Social Sciences and Humanities.

The exploration to the margins and beyond proposed by the theme has brought to the STI Conference an interesting array of new contributors from a variety of fields and geographies.

This year's conference had a record 382 registered participants from 40 different countries, including 23 European, 9 American, 4 Asia-Pacific, 4 Africa and Near East. About 26% of participants came from outside of Europe.

There were also many participants (17%) from organisations outside academia including governments (8%), businesses (5%), foundations (2%) and international organisations (2%). This is particularly important in a field that is practice-oriented.

The chapters of the proceedings attest to the breadth of issues discussed. Infrastructure, benchmarking and use of innovation indicators, societal impact and mission oriented-research, mobility and careers, social sciences and the humanities, participation and culture, gender, and altmetrics, among others.

We hope that the diversity of this Conference has fostered productive dialogues and synergistic ideas and made a contribution, small as it may be, to the development and use of indicators that, being more inclusive, will foster a more inclusive and fair world.

The organising committee

Jordi Molas-Gallart, Alejandra Boni, Elena Castro-Martínez, Ismael Rafols and Richard Woolley

PERIPHERIES, FRONTIERS AND BEYOND

This conference aims to stimulate reflection on the **challenges posed to S&T indicator development and use in geographical, cognitive or social spaces that are peripheral** or marginal to the centres of economic, scientific or technological activity. The focus is also on emerging areas of research and innovation that are inadequately described by existing, quantitative or qualitative indicators.

We propose to **identify, describe and analyse the problems that emerge in situations and spaces where indicators are used beyond their scope of validity**. The conference aims to offer an international platform to propose, and discuss, alternative approaches and indicators.

The conference will consider both weak (technical) and strong (socio-political) notions of periphery. The **weak notion understands peripheries as areas that are not adequately covered or targeted by current indicators**. The main concern here is the existence of indicator biases; the challenge lies in developing approaches and indicators that provide a more accurate or valid representation of science, technology and innovation activities.

The strong notion sees the periphery as composed by those having a lower status in an unequal or dependent relationship. It is therefore a relational concept in a situation that involves structural unequal access to resources. According to this view, peripheries tend to remain as such unless determined efforts to change their situation are undertaken and the use of indicators may contribute to build and sustain peripheral situations. The strong notion of periphery underlines the performative nature of indicators; that is, their capacity to shape reality.

The conference will consider various types of peripheral spaces. In the global economy, some geographical regions are often conceived as peripheral. Developing countries were long ago described as "the" periphery, but within every **geographical territory** we can also encounter peripheral zones (Southern European and Eastern European countries as peripheral to the European Union, poor regions are peripheral to the capital and richer regions within a country, etcetera). Specific problems also emerge in regions that undergo **socio-economic transitions** and are in need of implementing alternative (re)development strategies, in particular in relation to **sustainability**.

We can also refer to peripheral **social groups**: the disenfranchised, the poor, or perhaps the elderly. Research and innovation conducted in these spaces may require different types of indicators from the ones we are accustomed to use. There are also **cognitive peripheries**: areas of research that do not capture the attention of mainstream politicians and receive more limited resources. For example, many fields in the humanities could be considered a peripheral when compared to the mainstream natural sciences or engineering.

Each of these peripheries has their own knowledge generation and application systems and may be better analysed using **tailored indicators, some of which can be of a qualitative rather than quantitative nature.** However, analysts often face resource limitations to develop indicators tailored to the peculiarities of their context and are confronted with the potential use of conventional indicators –which are not fully suited to reflect these contexts. The use of such **indicators may result in inadequate analysis and unintended effects**.

The conference aims to be a platform to **reflect on the potential causes and effects of indicators usage in peripheral spaces**: in mobility and internationalisation, reduction of thematic diversity and alignment or misalignment with local societal needs.

COMMITTEES

CONFERENCE CHAIR

Jordi Molas-Gallart
Ingenio, CSIC-UPV, València

PROGRAMME BOARD

- Judith Sutz · Univ. de la República, Montevideo
- Tim Turpin · Univ. Western Sydney
- Sybille Hinze · DZHW, Berlin
- Robert Tijssen · CWTS, Univ. Leiden
- Jordi Molas-Gallart, Ingenio, CSIC-UPV, València
- Ismael Rafols · Ingenio, CSIC-UPV, València
- Elena Castro-Martínez · Ingenio, CSIC-UPV, València

LOCAL ORGANIZING COMMITTEE

- Alejandra Boni
- Elena Castro-Martínez
- Jordi Molas-Gallart
- Ismael Rafols
- Richard Woolley

SCIENTIFIC COMMITTEE

- Isidro F Aguillo · *IPP-CSIC, Spain*
- Juan Pablo Alperin · *Simon Fraser University , Canada*
- Rigas Arvanitis · *IFRIS, IRD, France*
- Joaquín M. Azagra · *INGENIO (CSIC-UPV), Spain*
- Rodolfo Barrere · *RICYT and Centro Redes, Argentina*
- Julieta Barrenechea · *Euskampus, Universidad del País Vasco, Spain*
- Katy Börner · *School of Informatics and Computing, Indiana University Bloomington, USA*
- Carolina Cañibano · *INGENIO (CSIC-UPV), Spain*
- Elena Castro-Martínez · *INGENIO (CSIC-UPV), Spain*
- Rodrigo Costas · *(CWTS) Leiden University, The Netherlands*
- Susan Cozzens, *Georgia Institute of Technology, Atlanta · USA*
- Laura Cruz Castro ·*IPP, CSIC, Spain*
- Pablo D'Este · *INGENIO (CSIC-UPV), Spain*
- Gabriela Dutrénit ·*Departamento de Producción Económica, Universidad Autónoma Metropolitana, México*
- Ernesto Fernández Polcuch · *Division of Science Policy and Capacity Building Natural Sciences Sector. UNESCO*
- Sergio Filho-Salles · *Departamento de Politica Cientifica e Tecnológica, State University of Campinas, Brazil*
- Fernando Galindo-Rueda · *OECD Directorate for Science, Technology and Industry, France*
- Fred Gault · *UNU-MERIT, Maastricth University, The Netherlands*
- Elea Giménez Toledo · *Instituto de Filosofía, CSIC, Spain*
- Wolfgang Glanzel, *ECOOM – KU Leuven, Belgium*
- Jochen Gläser · *Center for Technology and Society TU Berlin, Germany*
- Stefanie Haustein, *(EBSI) Université de Montréal, Canada*
- Sybille Hinze · *DZHW, Berlin, Germany*
- Hugo Hollanders · *UNU-MERIT, Maastricht University, The Netherlands*
- Wolfram Horstmann · University of Göttingen, *Germany*
- Erika Kraemer-Mbula · *IERI, Tshwane University of Technology, South Africa*
- Pablo Kreimer · *Universidad de Quilmes, Argentina*
- Frédérique Lang · *SPRU, University of Sussex, UK*
- Philippe Laredo · *IFRIS, Université Paris-Est, France*
- Vincent Larivière · *School of Library and Information Science, Université de Montreal, Canada*
- Thed van Leeuwen · *CWTS, Leiden Universit, The Netherlands*
- Loet Leydesdorff · *University of Amsterdam, The Netherlands*

- Sami Mahroum · *INSEAD, France*
- Luciana Marins · *UNESCO Institute of Statistics*
- Ramon Marrades · *Econcult, University of València, Spain*
- Catalina Martínez-García, *IPP, CSIC. Spain*
- Juan Mateos Garcia · *NESTA, UK*
- Niels Mejlgaard · *Danish Centre for Studies in Research and Research Policy, Aarhus University, Denmark*
- Julia Melkers · *Georgia Institute of Technology, USA*
- Jordi Molas-Gallart · *INGENIO (CSIC-UPV), Spain*
- Johan Mouton · *CREST, Stellenbosch University, South Africa*
- Cameron Neylon · *Curtin University, Perth, Australia*
- Ed Noyons · *(CWTS), Leiden University, The Netherlands*
- Julia Olmos · *Departament Direcció d'Empresas, University of València, Spain*
- Regina Pinto de Gusmao · *Brazilian Ministry of Science, Technology and Innovation (MCTI), Brazil*
- Guilherme Ary Plonski · F*aculdade de Economia, Administração e Contabilidade (FEA), Universidade de Sao Paulo, Brasil*
- Wolfgang Polt · *Joanneum Research, Austria*
- Alan Porter · *Georgia Tech, USA*
- Slavo Radosevic · *University College London, UK*
- Ismael Rafols · *INGENIO (CSIC-UPV), Spain*
- Emmanuela Reale · *IRCRES, CNR , Italy*
- Nicolas RobinsonGarcia · *INGENIO (CSIC-UPV), Spain*
- Daniele Rotolo · *SPRU, University of Sussex, UK*
- Ronald Rousseau, *KU Leuven, Belgium*
- Mónica Salazar · *InterAmerican Development Bank*
- Mabel Sánchez Barrioluengo · *ISPRA, European Commission, Italy*
- Thomas Scherngell · *Austrian Institute of Technology, Austria*
- Gunnar Siversten · *NIFU, Norway*
- Cassidy Sugimoto · *School of Informatics and Computing, Indiana University Bloomington, USA*
- Judith Sutz · *Universidad República, Uruguay*
- *Robert Tijssen* · CWTS, Univ. Leiden, The Netherlands
- Tim Turpin · *Univ. Western Sydney, Australia*
- Hebe Vessuri · *Instituto Venezolano de Investigaciones Científicas, Venezuela*
- Melanie Walker · *University Free State, South Africa*
- Matthew Wallace · *Canada Foundation for Innovation, Canada*
- Ludo Waltman *Leiden University, The Netherlands*
- Lili Wang · *UNU-MERIT, Maastricht University, The Netherlands*
- Inge van der Weijden · Leiden University, The Netherlands
- Richard Woolley · *INGENIO (CSIC-UPV), Spain*
- Paul Wouters · *CWTS, Leiden Universit, The Netherlands*
- Masaru Yarime · *Graduate School of Public Policy, University of Tokyo, Japan*
- Jon Mikel Zabala Iturriagagoitia · *University of Deusto, Spain*
- Zohreh Zahedi · *Leiden University, The Netherlands*
- Ping Zhou · *Department of Information Resources Management. Zhejiang University, China*
- Alesia Ann Zuccala · The Royal School of Library and Information Science. University of Copenhagen, Denmark

ORGANIZERS & SPONSORS

ORGANIZERS

PLATINUM SPONSORS

GOLD SPONSOR

SILVER SPONSOR

CONFERENCE PROGRAMME

PROGRAMME
SEPTEMBER 14

08.30-13.00 Registration

9.00-10.30 **OPENING SESSION AND 1ST PLENARY SESSION**

Wellcome to STI2016 - *Jordi Molas-Gallart* INGENIO (CSIC-UPV)

Wellcome to the Universitat Politècnica de València - *Francisco J. Mora Mas*, Rector

Keynote lecture: "The Power of Numbers: a critical review of MDG targets for Human Development" *Prof. Sakiko Fukuda-Parr* - The New School, New York, US

10.30-11.00 Coffee Break

11.00-12.30 PARALLEL SESSIONS 1

1.1 **SPECIAL TRACK 6** · Smart use of indicators for innovation policy

Chair: *Hugo Hollanders & Lili Wang*

Introduction – *Hugo Hollanders*

Innovation indicators: Towards a User's guide – *Michiko Iizuka, Hugo Hollanders*

Analyzing innovation policy indicators through a functional approach: the aeronautic industry case – *Carolina Resende Haddad, Mauricio Maldonado Uriona*

Assessing the performance of national innovation systems in Europe – *Jon Mikel Zabala-Iturriagagoitia*

1.2 Indicators, evaluation and policy

Chair: *Jesper Schneider*

Outlining an analytical framework for mapping research evaluation landscapes – *Fredrik Åström*

When the Brightest are not the Best– *Marco Valent*

The use of indicators and other evidence in two investment decisions of technology innovation – *Nuno F.F.G. Boavida*

1.3 Reward systems

The reward (eco)system of science: More than the sum of its parts? – A Special Fishbowl session – *Nadine Desrochers, Stefanie Haustein, Juan Pablo Alperin, Timothy D. Bowman, Adrián A. Díaz-Faes, Vincent Larivière, Philippe Mongeon, Adèle Paul-Hus, Anabel Quan-Haase, Elise Smith*

1.4 Knowledge Exchange

Chair: *Jan Youtie*

Knowledge integration through collaboration: building indicators using the Diversity/coherence and Proximity frameworks – *Frédérique Lang, Ismael Rafols, Michael Hopkins*

Using a network-based approach to identify interactions structure for innovation in a low-technology intensive sector – *Camille Aouinait*

Interdisciplinary and transdisciplinary institutions: do they constitute peripheries among cultures? – *Bianca Vienni, Ulli Vilsmaier*

“Putting in more than you take out”. Towards evaluating research based on its public (not private) contributions – *Paul Benneworth, Julia Olmos Peñuela, Elena Castro Martínez*

1.5 Careers and labour market

Chair: *Peter van den Besselaar*

Survey on the Labour Market Position of PhD Graduates – *Julia Heuritsch, Cathelijn Waaijer, Inge van der Weijden*

Beyond the indicators: formulation of the career strategies of scientists – *Eva Palinko*

Exploring predictors of scientific performance with decision tree analysis: The case of research excellence in early career mathematics – *Jonas Lindahl*

Stability and longevity in the publication careers of U.S. doctorate recipients – *Cathelijn Waaijer, Benoît Macaluso, Cassidy Sugimoto, Vincent Larivière*

1.6 **GENDER SPECIAL SESSION** on Gender in science: a periphery?

Chair: *Inge van der Weijden*

Gender equality and evaluation: do fields of science matter? – *Emanuela Reale, Antonio Zinilli*

Scientific and technological output of women and men – *Rainer Frietsch, Susanne Bührer, Patricia Helmich*

Gender and International Mobility of European Researchers – *Carolina Cañibano, Mary Frank Fox and F. Javier Otamendi*

Gender differences and the role of research grants – *Carter Bloch, Evanthia K. Schmidt*

Gender structured universities and their impact on mental health: a focus on PhD students in Flanders – *Katia Levecque & Frederik Anseel*

Gender differences in careers after receiving a personal grant – *Inge van der Weijden & Ingeborg Meijer*

1.7 Citation Impact

Chair: *Rogério Mugnaini*

Determinants of citation impact: A comparative analysis of the Global South versus the Global North – *Hugo Confraria*

Web of science coverage and scientific performance of Central and Eastern European countries – *Adam Ploszaj, Agnieszka Olechnicka*

Does size matter? An investigation of how department size and other organizational variables influence on publication productivity and citation impact – *Dag W Aksnes, Kristoffer Rørstad, Fredrik N Piro*

Do usage and scientific collaboration associate with citation impact? – *Pei-Shan Chi, Wolfgang Glänzel*

12:30-14:30 **POSTER SESSION 1** (see Appendix 1)

13:00-14:30 Lunch

14:30-16:00 PARALLEL SESSIONS 2

2.1 Resource distribution and research contents
Chair: *Jochen Gläser*

Unveiling Research Agendas: a study of the influences on research problem selection among academic researchers – *Mariela Bianco, Judith Sutz*

"If we come out with the wrong answer that may affect investments": Exploring how evaluators were influenced by political considerations during the assessment of societal impact – *Gabrielle Samuel, Gemma Elizabeth Derrick*

Must Metrics Serve the Audit Society? Addressing Marketization in Open Access Publishing and Humanities Analytics – *Christopher Newfield, Christopher Muellerleile*

2.2 **SPECIAL SESSION** on multiplying methods in the field of research evaluation
Chair: *Inge van der Weijden*

Introduction & recap: *Gemma Derrick*

Provocation: *Paul Wouters*

Demonstrations in pairs:

Wolfgang Kaltenbrunner & Michael Ochsner

Ingeborg Meijer & Carolina Cañibano

Rogerio Mugnaini & Nadine Desrochers

Next steps: *Gemma Derrick, Jordi Molas-Gallart and Sarah de Rijcke + Irene Ramos-Vielba*

2.3 **SPECIAL TRACK 6** · Smart use of indicators for innovation policy
Chair: *Hugo Hollanders & Lili Wang*

Evidence-based policy learning: the case of the Research Excellence Indicator
Sjoerd Hardeman, Daniel Vertesy

Who sets up the bridge? Tracking scientific collaborations between China and the European Union – *Lili Wang*

A case study about the Colombian Observatory of Science and Technology: between context relevant and internationally comparable indicators – *Mónica Salazar*

2.4 **SPECIAL TRACK 1** · Data infrastructure and data quality for evolving research metrics

Chair: *Chris Keene*

Introduction – *Chris Keene*

On the Peripheries of Scholarly Infrastructure: A look at the Journals Using Open Journal Systems – *Juan Pablo Alperin, Kevin Stranack, Alex Garnett*

Why researchers publish in journals not indexed in mainstream databases: training, bridging and gap-filling – *Diego Chavarro, Puay Tang, Ismael Rafols*

Identifying Sources of Scientific Knowledge: classifying non-source items in the WoS – *Clara Calero Medina*

2.5 Careers and labour market

Chair: *Pablo D'Este*

Developing research career indicators using open data: the RISIS infrastructure – *Carolina Cañibano, Richard Woolley, Eric Iversen, Sybille Hinze, Stefan Hornbostel, Jakob Tesch*

On the extent of researcher mobility and indicators for mobility – *Stina Gerdes Barriere*

Progress on mobility and instability of research personnel in Japan: scientometrics on a job-posting database for monitoring the academic job market – *Hirotaka Kawashima, Yasuhiro Yamashita*

National and international scientific elites: an analysis of Chinese scholars
Fei Shu, Vincent Larivière, Charles-Antoine Julien

2.6 Gender

Chair: *Monica Gaughan*

What drives the gender gap in STEM? The SAGA Science, Technology and Innovation Gender Objectives List (STI GOL) as a new approach to linking indicators to STI policies – *Ernesto Fernández Polcuch, Martin Schaaper, Alessandro Bello*

Picking the best publications to showcase graduate courses: Do institutional mechanisms reinforce gender differences? – *Jacqueline Leta, Guillaume Cabanac*

What factors influence scientific and technological output: A comparison of Thailand and Malaysia – *Catherine Beaudry, Carl St-Pierre*

2.7 Citation Impact

Chair: *Wolfgang Glänzel*

An approach for the condensed presentation of intuitive citation impact metrics which remain reliable with very few publications – *David Campbell, Chantale Tippett, Gregoire Cote, Guillaume Roberge, Eric Archambault*

A comparison of average-based, percentile rank, and other citation impact indicators – *Pedro Albarrán, Javier Ruiz-Castillo*

How does the scientific progress in developing countries affect bibliometric impact measures of developed countries? A counterfactual case study on China– *Stephan Stahlschmidt, Sybille Hinze*

The returns to scientific specialization – *Orion Penner, Gaétan de Rassenfosse*

16:00-16:30 Coffee Break

16:30-18:00 **2ND PLENARY SESSION** [Sponsored by RISIS]

Roundtable: Infrastructures for Inclusive and Open Science and RISIS presentation

Chair: *Ismael Rafols*

Éric Archambault, Science-Metrix, Montréal, Canada
Chris Keene, JISC, UK
Valentin Bogorov, Thomson-Reuters, Moscow, Russia
Abel Packer Scielo, Sao Paulo, Brazil
Hebe Vessuri, IVIC, Venezuela
Emanuela Reale, IRCRES, CNR, Italy

18:30-19:00 Transfer to the cocktail site

19:00-21:00 **WELCOME COCKTAIL** [SANT MIQUEL DELS REIS]

09.00-10.30 PARALLEL SESSIONS 3

3.1 **SPECIAL TRACK 3** · Measuring diverse research "qualities": indicators of societal impact, engagement, participation, and local relevance

Chair: *Judith Sutz*

Introduction – *Judith Sutz*

'Productive interactions' for societal impact: developing a research information system for agriculture (RIS-Agric) at Stellenbosch University, South Africa – *Nelius Boshoff, Harrie Esterhuyse*

Publication patterns in research underpinning impact in REF2014 – *Jonathan Adams*

3.2 University-Industry relations

Chair: *Puay Tang*

Measuring macro-level effects of the global economic recession on university-industry research cooperation – *Joaquín M. Azagra-Caro*

3.3 **SPECIAL TRACK 5** · Social sciences and the humanities

Chair: *Thed van Leeuwen*

Introduction – *Thed N. van Leeuwen*

Indicators for research performance in the humanities? The scholars' view on research quality and indicators – *Michael Ochsner, Sven E. Hug*

Quality criteria and indicators for research in theology – What to do with quantitative measures? – *Silvia Martens, Wolfgang Schatz*

3.4 **SPECIAL TRACK 1** · Data infrastructure and data quality for evolving research metrics

Chair: *Chris Keene*

Data quality and consistency in scopus and Web of science in their indexing of Czech Journals – *Pavel Mika, Jakub Szarzec, Gunnar Sivertsen*

Missing citations due to exact reference matching: analysis of a random sample from WoS. Are publications from peripheral countries disadvantaged? *Paul Donner*

Funding acknowledgements in the Web of science: inconsistencies in data collection and standardization of funding organizations – *Jeroen van Honk, Rodrigo Costas, Clara Calero-Medina*

Open data in global environmental change: findings from the community – *Birgit Schmidt*

3.5 **SPECIAL TRACK 2** · International benchmarking of innovation: challenges and adequacy for developing and developed regions

Chair: *Luciana Marins*

Introduction – *Luciana Marins*

The impact of methodology in innovation measurement – *Espen Solberg, Lars Wilhelmsen, Markus Bugge*

A critical assessment of the quality and validity of composite indicators of innovation – *Daniel Vertesy*

Innovation strategies in Latin American firms – *Fernando Vargas*

3.6 Gender

Chair: *Jacqueline Leta*

Identifying the gender dimension in research content – *Chantale Tippett, David Campbell, Bastien St. Louis Lalonde, Eric Archambault, Julie Callaert, Katerina Mantouvalou, Lucy Arora*

Gender differences in synchronous and diachronous self-citations – *Gita Ghiasi, Vincent Larivière, Cassidy Sugimoto*

Mapping the author gender-distribution of disease-specific medical research *Jens Peter Andersen, Jesper Wiborg Schneider, Mathias Wullum Nielsen*

Indicators for constructing scientific excellence: "Independence" in the ERC starting grant – *Helene Schiffbaenker, Florian Holzinger*

3.7 Citation Impact

Chair: *Erjia Yan*

A comparison of the Web of science with publication-level classification systems of science – *Antonio Perianes-Rodriguez, Javier Ruiz-Castillo*

Ranking journals using social choice theory methods: a novel approach in bibliometrics – *Fuad Aleskerov, Vladimir Pislyakov, Andrey Subochev*

The performance and trend of China's academic disciplines from 2006 to 2014 – *Zhigang Hu*

Comparing absolute and normalized indicators in scientific collaboration: a study in Environmental Science in Latin America – *Maria Cláudia Cabrini Grácio, Ely Francina Tannuri de Oliveira*

10.30-11.00 Coffee break

11.00-12.30 PARALLEL SESSIONS 4

4.1 **SPECIAL GLOBELICS SESSION** on Lessons learned for priority setting and indicators relevant to the impact of research programmes in Europe and Emerging Economies. An evidence-based debate between the research and the policy-shaping community

Chair: *Yannis Caloghirou, Nicholas Vonortas*

Thirty years of European Collaboration in Research and Development: Policy-driven Research Networking and the presence of new knowledge-intensive entrepreneurial ventures – *Yannis Caloghirou, Aimilia Protogerou and Evangelos Siokas*

STI Indicators for Emerging Economies: Experiences from Chile, Brazil and Peru *Adriana Bin, Sergio Salles-Filho, Ana Maria Carneiro, Nicholas Vonortas, Juan Ernesto Sepulveda Alonso and Paula Felicio Drummond de Castro*

Use of indicators for research and policy impact evaluation: evidence from Russia – *Konstantin Fursov and Stanislav Zaichenko*

4.2 **SPECIAL TRACK 3** · Measuring diverse research "qualities": indicators of societal impact, engagement, participation, and local relevance

Chair: *Judith Sutz*

Societal impact metrics for non-patentable research in dentistry – *Diana Hicks, Kim Isett, Julia Melkers, Le Song, Rakshit Trivedi*

The Evolution of Scientific Trajectories in Rice: Mapping the Relation between Research and Societal Priorities – *Tommaso Ciarli, Ismael Rafols*

Research Quality Plus (RQ+) A Holistic Approach to Evaluating Research *Robert McLean, Osvaldo Feinstein*

4.3 **SPECIAL TRACK 5** · Social sciences and the humanities

Chair: *Thed van Leeuwen*

Social Impact Open Repository (SIOR). Transforming the peripheral space of social impact of research – *Mar Joanpere, Elvira Samano*

Je veux bien, mais me citerez-vous? On publication language strategies in an anglicized research landscape – *Nadine Desrochers, Vincent Larivière*

Effects of performance-based research funding on publication patterns in the social sciences and humanities – *Raf Guns, Tim Engels*

4.4 Indicators and infrastructure

Chair: *Eric Archambault*

Examining data access and use in science – *Erjia Yan, Mengnan Zhao*

SMS: a linked open data infrastructure for science and innovation studies *Peter Van den Besselaar, Ali Khalili, Al Idrissou, Antonis Loizou, Stefan Schlobach, Frank Van Harmelen*

Data Citation Policies of Data Providers within the scope of Longitudinal Studies in Life Course Research – *Anke Reinhold, Marc Rittberger, Nadine Mahrholz*

Stepping up Information Infrastructures and Statistical Reporting-Monitoring the German Excellence Initiative – *Anke Reinhardt*

4.5 **SPECIAL TRACK 2** · International benchmarking of innovation: challenges and adequacy for developing and developed regions

Chair: *Luciana Marins*

Innovation dynamics of Salvadoran agro-food industry from an evolutionary perspective – *Elías Humberto Peraza Castaneda, Guillermo Aleixandre Mendizábal*

Elucidate Innovation Performance of Technology-driven Mergers and Acquisitions – *Lu Huang, Kangrui Wang, Huizhu Yu, Lining Shang, Liliana Mitkova*

4.6 Society, participation and culture

Chair: *Julia Melkers*

Operationalizing RRI: Relational Quality Assessment & Management Model for Research and Innovation Networks (REQUANET) – *Julieta Barrenechea, Andoni Ibarra*

What knowledge counts? Insights from an action research project using participatory video with grassroots innovation experiences – *Alejandra Boni, Monique Leivas, Alba Talón, Teresa De la Fuente, Victoria Pellicer-Sifres, Sergio Belda-Miquel, Aurora López-Fogués, Begoña Arias*

A proposal for measurement of science and innovation culture – *Asako Okamura*

4.7 Individual Performance

Chair: *María Bordons*

Information sources – information targets: evaluative aspects of the scientists' publication strategies – *Wolfgang Glanzel, Pei-Shan Chi, Christian Gumpenberger, Juan Gorraiz*

The Effect of Holding a Research Chair on Scientists' Impact – *Seyed Reza Mirnezami, Catherine Beaudry*

Public-private collaboration and scientific impact: an analysis at the level of the individual researcher – *Carter Bloch, Thomas K. Ryan, Jens Peter Andersen*

4.8 Funding and EU collaboration

Chair: *Philippe Larédo*

Examining to What Extent Does the Source of Funding Matter for Scientific Impact. A Case Study of Danish EU FP7 Funded Projects – *Thomas Kjeldager Ryan, Jesper Wiborg Schneider*

The Determinants of National Funding in Trans-national Joint Research: Exploring the Proximity Dimensions – *Emanuela Reale, Andrea Orazio Spinello, Antonio Zinilli*

Beyond funding: What can acknowledgements reveal about credit distribution in science? – *Adèle Paul-Hus, Adrián A. Díaz-Faes, Nadine Desrochers, Rodrigo Costas, Maxime Sainte-Marie, Benoît Macaluso and Vincent Larivière*

Allocating organisational level funding on the basis of Research Performance Based assessments, a comparative analysis of the EU Member States in international perspective – *Koen Jonkers, Thomas Zacharewicz, Benedetto Lepori, Emanuela Reale*

12.30-13.30 **3RD PLENARY SESSION** [Sponsored by IFRIS]

Roundtable: Global collaboration networks: flat world or centre-periphery structure?

Chair: *Richard Woolley*

Jonathan Adams, Digital Science, London, UK
Rigas Arvanitis, Director of IFRIS, IRD, Paris, France
Sami Mahroum, INSEAD Innovation and Policy Initiative, Abu Dhabi, United Arab Emirates
Mónica Salazar, InterAmerican Develoment Bank, Bogota, Colombia

13.30-14.30 Lunch

14:30-16:00 PARALLEL SESSIONS 5

5.1 **SPECIAL TRACK 5** · Social sciences and the humanities

Chair: *Thed van Leeuwen*

Developing appropriate methods and indicators for evaluation of research in the social sciences and humanities. Presentation of a new COST Action
Gunnar Sivertsen, Ioana Galleron

5.2 **SPECIAL TRACK 3** · Measuring diverse research "qualities": indicators of societal impact, engagement, participation, and local relevance

Chair: *Judith Sutz*

Impact of Research on Development in Cameroon: convergence between supply and research needs in the food sector – *Minkoua Jules René, Ludovic Temple*

Monitoring the Evolution and Benefits of Responsible Research and Innovation (MoRRI) – a preliminary framework for measuring RRI dimensions – *Niels Mejlgaard, Susanne Buehrer, Erich Griessler, Ralf Lindner, Nikos Maroulis, Ingeborg Meijer, Viola Peter, Ismael Rafols, Tine Ravn, Jack Stilgoe, Lena Tsipouri, Richard Woolley, Angela Wroblewski*

"All this grassroots, real life knowledge": Comparing perceived with realised concerns of including non-academic evaluators in societal impact assessment – *Gemma Derrick, Gabrielle Samuel*

5.3 **SPECIAL SESSION** on Predicting STEM Career Success by STI Knowledge Utilization Patterns

Chair: *Barry Bozeman, Jan Youtie*

Career Impacts of Cosmopolitan Collaboration – *Barry Bozeman, Monica Gaughan*

Bounded Collaboration and Changing Core-Periphery Relationships in Sino-Russian Scientific Co-Production – *Abdullah Gök, Maria Karaulova, Philip Shapira*

Going home: why do non-US citizens with US Ph.D. degrees return home? *Stuart Bretschneider*

The credibility of policy reporting across learning disciplines – *Jan Youtie*

5.4 **SPECIAL SESSION** on Performance indicators for areas of innovation: international perspective

Chair: *Guilherme Ary Plonski*

A case study of Be'er Sheva Advanced Technology Park (ATP) in Israel – *Daphne Getz, Eliezer Shein*

Porto Digital: an area of innovation as a lever to transform Recife in Brazil *Guilherme Ary Plonski, Désirée M. Zouain*

The influence of Science and Technology parks in Spain – *Andres Barge-Gil, Aurelia Modrego Rico*

5.5 Mission Oriented Research Health

Chair: *Sandro Mendonça*

Using novel computer-assisted linguistic analysis techniques to assess the timeliness and impact of FP7 Health's research – *Vilius Stanciauskas*

Professional impact – *Gustaf Nelhans*

Technology push / market pull indicators in healthcare – *Irina Efimenko, Vladimir Khoroshevsky, Ed Noyons, Evgeny Nochevkin*

Mapping the networks of cancer research in Portugal (1990-2015): initial results – *Oriana Rainho Brás, Jean-Philippe Cointet, João Arriscado Nunes, Leonor David, Alberto Cambrosio*

5.6 Text analysis

Chair: *Stefan Hornbostel*

Breakout discoveries in science: what do they have in common? – *Jos Winnink, Robert J.W. Tijssen, Anthony F.J. van Raan*

From university research to innovation – detecting knowledge transfer via text mining – *Sabrina Larissa Woltmann, Line H. Clemmensen, Lars Alkærsig*

Predicting panel scores by linguistic analysis– *Peter Van den Besselaar*

5.7 Altmetrics

Chair: *Stefanie Haustein*

Article-level metrics and the periphery: an exploration of articles by Brazilian authors – *Iara Vidal Pereira de Souza, Fabio Castro Gouveia*

Can we use altmetrics at the institutional level? A case study analysing the coverage by research areas of four Spanish universities – *Daniel Torres-Salinas, Nicolas Robinson-Garcia, Evaristo Jiménez-Contreras*

Enhancing methodology of altmetrics studies by exploring social media metrics for Economic and Business Studies journals – *Kaltrina Nuredini, Isabella Peters*

Comparative study of Colombian Researchers according to data from Google Scholar, ResearchGate and the National System for Measurement Science (Colciencias) – *Isidro F Aguillo, Alejandro Uribe-Tirado, Wilson López*

16:00-16:30 Coffee Break

16:30-18:00 PARALLEL SESSIONS 6

6.1 Altmetrics – PANEL

Roundtable: Next-generation metrics: responsible metrics & evaluation for open science

James Wilsdon, Judit Bar-Ilan, Isabella Peters, Paul Wouters

Chair: *James Wilsdon*

6.2 Geography and performance

Chair: *Michael J Kahn*

Indicators of the knowledge based society: Comparison between European and Latin American countries – *Daniel Villavicencio*

Measuring cross-border regional STI integration – *Teemu Makkonen*

From emerging country to a leading role in the scientific and technological field? analysis of the internationalization of Brazil – *Claudia Daniele de Souza, Daniela De Filippo, Elias Sanz Casado*

6.3 **SPECIAL TRACK 5** – Social sciences and the humanities

Chair: *Thed van Leeuwen*

Clashing Conventions? Exploring Human Resource Management in the Cleavage Between Academic Field Traditions and New Institutional Rules. Quantitative and Qualitative Insights from the Field of Communication and Media Studies in Switzerland – *Alexander Buhmann*

A bibliometric indicator with a balanced representation of all fields – *Gunnar Sivertsen*

Measuring research in humanities and social sciences: information from a new Italian data infrastructure – *Marco Malgarini, Tindaro Cicero*

Trends and developments in multi-authorship in five social science disciplines from 1991 to 2014 – *Sabrina Jasmin Mayer*

6.4 **SPECIAL TRACK 4** · Collaborations, mobility and internationalization

Chair: *Rigas Arvanitis*

Introduction: *Rigas Arvanitis*

Mobility in the academic careers at the flemish universities – Results from the human resources in research database – *Noëmi Frea Debacker, Karen Vandevelde*

Gatekeeping African studies: A preliminary insight on what do editorial boards indicate about the nature and structure of research brokerage – *Sandro Mendonça*

6.5 Mission Oriented Research – Health

Chair: *Matthew Wallace*

Access to global health research. Prevalence and cost of gold and hybrid open access – *Stefanie Haustein, Elise Smith, Philippe Mongeon, Fei Shu, Vincent Larivière*

Scientific research on diseases: the distinct profile of developed and developing countries – *Alfredo Yegros*

Biodiversity sustainability of phytomedicine research:a 3-dimensions analysis around the North-South divide – *Philippe Gorry*

In Re the academic cartography of sugar sweetened beverages: scientific and technical information, interdisciplinarity, and legal academia – *Lexi C. White*

6.6 Project and programme assessment

Chair: *Diana Hicks*

An assessment of EU-funded research projects: innovators and their innovative potential – *Daniel Nepelski, Vincent Van Roy, Eoghan O'neill*

Evaluating the impact of public space investments with limited time and funds: (methodological) lessons from a Swiss case study – *Franz Barjak*

Researchers and institutions in the periphery: challenges in measuring research capacity for geographically specific programs in the U.S – *Julia Melkers*

Assessing marine biotechnology research centres in peripheral regions: developing global and local STI indicators – *Antoine Schoen, Douglas Robinson*

6.7 Measuring Innovation

Chair: *Joaquín M. Azagra-Caro*

Baseline of indicators for R&D and Innovation in ICT: a tool for decision-making, design and monitoring of public policies – *Henry Mora Holguín, Diana Lucio-Arias, Sandra Zárate, Nayibe Castro, Clara Pardo*

Measuring originality: common patterns of invention in research and technology organizations – *David Li Tang, Erica Wiseman, Tamara Keating, Jean Archambeault*

Linking international trademark databases to inform IP research and policy *Stephen Michael Petrie*

Detecting emerging trends and country specializations in energy efficiency – *Daniela De Filippo, Andres Pandiella-Dominique, Elba Mauleon*

18:00-19:30 ENID General Assembly

18:00-19:30 **SPECIAL SESSION** "Ciencia, Tecnología, Sociedad e Innovación ¿Medimos lo que debemos?¿Medimos bien?"
Organised and sponsored by the Cátedra CTSi (OEI-Junta de Andalucía)
(This special session is in Spanish)

Chair: *Manuel Torralbo* Junta de Andalucía, Spain

Judith Sutz – Universidad de la República, Uruguay
Hebe Vessuri – CONICET, Argentina
José Navarrete – Junta de Andalucía, Spain

20:00-22:00 **CONFERENCE DINNER** (Hotel Astoria)

09.00-10.30 **4th PLENARY SESSION**

Conference "The deep structure of STI indicators: Contextual knowledge and scientometrics"

Chair: *Philippe Laredo*

Keynote speaker: *Prof. Johann Mouton* – Stellenbosch University, South Africa

10.30-11.00 Coffee Break

11.00-12.30 PARALLEL SESSIONS 7

7.1 Innovation in Government

Roundtable: **SPECIAL SESSION** on Measuring Innovation in Government – *Anthony Arundel, Carter Bloch, Ilka Lakaniemi, Sami Mahroum*

Chair: *Sami Mahroum*

7.2 Mission Oriented Research – Agriculture

SPECIAL PANEL on Metrics and Agricultural Science measuring Multidisciplinary and Applied Research – *Vanessa Méry, Hugo Besemer, Ellen Fest, Soizic Messiaen*

Chair: *Ilkay Holt*

7.3 **SPECIAL TRACK 5** · Social sciences and the humanities

Chair: *Thed van Leeuwen*

ERIH PLUS – Making the SSH visible, searchable and available – *Gry Ane Vikanes Lavik, Gunnar Sivertsen*

Indexed University presses: overlap and geographical distribution in five book assessment databases – *Jorge Mañana-Rodríguez, Elea Giménez-Toledo*

East-African Social Sciences and Humanities Publishing – A Handmade Bibliometrics Approach – *Nora Schmidt*

Alphabetical co-authorship in the social sciences and humanities: evidence from a comprehensive local database – *Raf Guns*

7.4 **SPECIAL TRACK 4** · Collaborations, mobility and internationalization

Chair: *Rigas Arvanitis*

Scientific mobility of Early Career Researchers in Spain and The Netherlands through their publications – *Nicolas Robinson-Garcia, Carolina Cañibano, Richard Woolley, Rodrigo Costas*

The network of international student mobility – *Eva Maria Voegtle, Michael Windzio*

Big Science, co-publication and collaboration: getting to the core – *Michael J Kahn*

Autonomy vs. dependency of scientific collaboration in scientific performance
Zaida Chinchilla-Rodríguez, Sandra Miguel, Antonio Perianes-Rodríguez, María-Antonia Ovalle-Perandones, Carlos Olmeda-Gómez

7.5 Culture and engagement

Chair: *Bianca Vienni*

Scientific culture in Colombia. A proposal of an indicator system for science, technology and innovation – *Clara Pardo, William Alfonso*

How user-innovators can be identified? Evidence collected from the analysis of practices – *Konstantin Fursov*

Assessing youth engagement with a collaborative Index – *Ramón Marrades*

7.6 Networks

Chair: *Ludo Waltman*

Networks dynamics in the case of emerging technologies – *Daniele Rotolo*

Using network centrality measures to improve national journal classification lists – *A. Zuccala, N. Robinson-Garcia, R. Repiso, D. Torres-Salinas.*

Bridging centrality: A new indicator to measure the positioning of actors in R&D networks – *Thomas Scherngell, Laurent Berge, Iris Wanzenböck*

Network heterogeneity in an undirected network – *Xiaojun Hu, Loet Leydesdorff, Ronald Rousseau*

7.7 Altmetrics

Chair: *Juan Pablo Alperín*

A Systematic Identification of Scientists on Twitter – *Qing Ke, Yong-Yeol Ahn, Cassidy R. Sugimoto*

Do Mendeley reader counts reflect the scholarly impact of conference papers? A comparison between ComputerScience and Engineering field. – *Kuku Joseph Aduku, Mike Thelwall, Kayvan Kousha*

Currencies of science: discussing disciplinary “exchange rates” for citations and Mendeley readership – *Rodrigo Costas, Antonio Perianes-Rodríguez, Javier Ruiz-Castillo*

SSH & the City. A Network Approach for Tracing the Societal Contribution of the Social Sciences and Humanities for Local Development – *Nicolas Robinson-Garcia, Thed N. van Leeuwen, Ismael Rafols*

12.30-14.30 POSTER SESION 2 (see Appendix 2)

13:00-14:30 Lunch

14:30-16:00 PARALLEL SESSIONS 8

8.1 Indicators' use and effects

Chair: *Paul Wouters*

Why DORA does not stand a chance in the biosciences – *Jochen Gläser*

Are institutional missions aligned with journal-based or document-based disciplinary structures? – *Richard Klavans, Kevin Boyack*

Science policy through stimulating scholarly output Does is work? – *Peter Van den Besselaar*

The need for contextualized scientometric analysis: An opinion paper – *Ludo Waltman*

8.2 National systems in the periphery

Chair: *Daniel Villavicencio*

Measuring internationality without bias against the periphery – *Valeria Aman*

Indicators on measuring technology convergence worldwide – *Chunjuan Luan*

Development on the Periphery: monitoring science, technology and innovation for sustainable development among Pacific Island Countries – *Tim Turpin, Ranasinghe Wasantha Amaradasa*

Fake Academic Degrees as an Indicator for Severe Reputation Crisis in the Scientific Community - *Andrey Rostovtsev, Alexander Kostinskiy*

8.3 **SPECIAL TRACK 5** · Social sciences and the humanities

Chair: *Thed van Leeuwen*

A SPECIAL DEBATE on Aligning research assessment in the Humanities to the national Standard Evaluation Protocol Challenges and developments in the Dutch research landscape – *Ad Prins, Jack Spaapen, Frank van Vree*

8.4 **SPECIAL TRACK 4** · Collaborations, mobility and internationalization

Chair: *Rigas Arvanitis*

The world network of scientific collaborations between cities: domestic or international dynamics? – *Marion Maisonobe, Denis Eckert, Michel Grossetti, Laurent Jégou, Béatrice Milard*

Trends in the inter-regional and international research collaboration of the PRC's regions: 2000-2015 – *Marc Luwel, Erik van Wijk, Lambertus (Bert) J van der Wurff, Lili Wang*

Iran's scientific dominance and the emergence of South-East Asian countries in the Arab Gulf Region – *Henk F. Moed*

How international is internationally collaborated research? A bibliometric study of Russian surname holder collaboration networks – *Maria Karaulova, Abdullah Gök and Philip Shapira*

8.5 Mission-Oriented Research-Health

Chair: *Tommaso Ciarli*

Network analysis to support research management: evidence from the Fiocruz Observatory in Science, Technology and Innovation in Health – *Bruna de Paula Fonseca e Fonseca Fonseca, Ricardo Barros Sampaio, Marcus Vinicius Pereira da Silva, Paula Xavier dos Santos*

Partial alphabetical authorship in medical research: an exploratory analysis
Philippe Mongeon, Elise Smith, Bruno Joyal, Vincent Larivière

The bibliometric behaviour of an expanding specialisation of medical research
Jonathan Levitt, Mike Thelwall

8.6 **SPECIAL SESSION** · Scientific Culture Measures. Challenges and New Perspectives

Presentation / Introduction to the topic: What is scientific culture and what is not? – *José Antonio López Cerezo*

What does it mean to be scientifically literate? – *Belén Laspra*

New tools and indicators to measure scientific culture – *Ana Muñoz van den Eynde*

New cultural factors influencing the innovation measures – *María Cornejo Cañamares*

8.7 Altmetrics

Chair: *Rodrigo Costas*

Comparing the characteristics of highly cited titles and highly alted titles
Fereshteh Didegah, Timothy D. Bowman, Sarah Bowman, James Hartley

What makes papers visible on social media? An analysis of various document characteristics – *Zohreh Zahedi, Rodrigo Costas, Vincent Lariviere, Stefanie Haustein*

Normalization of Mendeley reader impact on the reader- and paper-side
Robin Haunschild, Lutz Bornmann

16:00-16:30 Coffee Break

16:30-18:00 **5TH PLENARY SESSION** [Sponsored by Thomson Reuters]

Roundtable: Use of indicators in policy and inclusive metrics

Chair: *Jordi Molas-Gallart*

Richard Deiss, Directorate General for Research and Innovation, European Commission

Diana Hicks, Georgia Tech, Atlanta, US

Slavo Radosevic, UCL, London, UK

Judith Sutz, President of Globelics & Univ. de la República, Montevideo, Uruguay

18:15-19:30 STI2016 Fringe

OPEN SESSION on local examples of participatory research

Video presentations

A roundtable on quality criteria and indicators for Participatory Action Research

Sandra Boni, Ramon Marrades and local Valencian activists

18:30-21:30 **CLOSING COCKTAIL & MUSIC**

12:30-14:30 POSTER SESSION 1

INDICATORS, ASSESSMENT, FUNDING AND INNOVATION

The research activity index at the Universitat Politècnica de València (IAIP): How an institution can complement national regulation on the productivity of university professors in research and teaching activities. *Conejero, J. Alberto; Capilla, José; Sánchez-Ruiz, Luis; Amigó, Vicente; Blasco, Agustín; Botti, Vicent; Cano, Juan; Capmany, José; Chiralt, Amparo*

Bibliometric indicators and activity scores for digital scholars. *Mikki, Susanne; Zygmuntowska, Marta*

Mapping scientific controversy in Twitter: the Maya city hoax. *Denia, Elena*

Visibility and Impact of Research Data Sets in the Life Sciences supported by a Novel Software Infrastructure. *Kramer, Claudia; Jung, Nicole; Tremouilhac, Pierre*

Changes in Scholars' Scientific Knowledge Production Shaped by Bibliometric Measures in Taiwan. *Peng, Ming-Te*

Purpose-oriented metrics to assess researcher quality; *Duarte, Kedma; Weber, Rosina; Pacheco, Roberto C.S.*

On the relationship between research topics and scientific impact: a study of edible animal research. *Castelló-Cogollos, Lourdes; Aleixandre-Benavent, Rafael; d'Este, Pablo; Rafols, Ismael*

Evaluation of grants schemes in the context of the national research system based on the publication count and citation data: the grants of the Latvian Council of Science. *Kokorevics, Arnis*

New approaches to monitor and evaluate Science, Technology and Innovation in health: a pilot study on the Zika virus. *Santos, Paula; Feltrin, Rebeca; Fonseca e Fonseca, Bruna; Barros, Ricardo; Reis, Juliana Gonçalves; Barreto, Jorge; Martins, Fatima; Barreto, Maurício; Lima, Nísia Trindade*

Issues relating to a Brazilian model of graduate courses evaluation: the CAPES system. *Vogel, Michely J.M.; Kobashi, Nair Y.*

Performance Based Funding and Researchers' Grant Application Strategies. *Johann, David; Neufeld, Jörg*

Impact of research evaluation modes of public research funding on the development of research fields and groups in Estonia. *Valdmaa, Kaija; Tõnurist, Piret*

The More Funding Sources, the More Citations? The Feasibility Study of Design on "Funding Diversity Indicator". *Chen, Carey Ming-Li*

12:30-14:30 POSTER SESSION 1

INDICATORS, ASSESSMENT, FUNDING AND INNOVATION

Accuracy and completeness of funding data in the Web of Science. *Álvarez-Bornstein, Belén; Morillo, Fernanda; Bordons, María*

Patent indicators for the Spanish nanotechnology domain. *Jürgens, Björn; Herrero-Solana, Víctor*

Best-Practice Benchmarking for Israel: The SNI Scorecard – A Multidimensional Perspective. *Maital, Shlomo; Buchnik, Tsipy; Getz, Daphne*

Measuring Global Innovation Activities with Article Visiting Geographical Data. *Wang, Xianwen; Fang, Zhichao; Yang, Yang; Wang, Hongyin; Hu, Zhigang*

Public scientists contributing to local literary fiction. An exploratory analysis. *Azagra-Caro, Joaquín M.; Fernández-Mesa, Anabel; Robinson-Garcia, Nicolas*

Does collaboration facilitate the performance of enterprise innovation? *Lv, Qi; Zhu, Donghua; Huang, Ying; Mitkova, Liliana; Wang, Xuefeng; Ogsuz, Gizem*

Structural Analysis of Redundancy Influence of Local Regions in Renewable Energy R&D Projects in Europe. *Larruscain-Sarasola, Jaso; Rio Belver, Rosa María; Garechana, Gaizka*

The discrepancy of patent citation behavior between examiners and inventors: a citation network analysis. *Huang, Ying; Zhu, Donghua; Lv, Qi; Porter, Alan L.; Wang, Xuefeng*

How Does Technology Transfer from Universities to Market in China? An Empirical Analysis Based on Invention Patent Assignment. *Yang, Yang; Ding, Kun; Zhang, Chunbo; Sun, Xiaoling; Hu, Zhigang*

Large Scale Disambiguation of Scientific References in Patent Databases. *Zhao, Kangran; Caron, Emiel; Guner, Stanisław*

12:30-14:30 POSTER SESSION 2

DATA CHARACTERISATION, CLASSIFICATION, VISUALISATION AND INDICATOR DESIGN

Quantifying and visualizing different types of scientific collaboration in Nanoscience and Nanotechnology field. *Chinchilla-Rodríguez, Zaida; Miguel, Sandra; Perianes-Rodríguez, Antonio*

Internal Migration of Scientists in Russia and the USA: the Case of Physicists. *Dyachenko, Ekaterina*

The Global Research Identifier Database GRID – Persistent IDs for the World's Research Organisations. *Szomszor, Martin; Mori, Andres*

Differential Effects of Scopus vs. Web of Science on University Rankings: A Case Study of German Universities. *Horstmann, Wolfram; Schmidt, Birgit*

On the normalization of citation impact based on the Essential Science Indicators classification of Thomson Reuters. *Baranova, Olga; Peris, Alfred*

Rock around the clock? Exploring scholars' downloading patterns. *Cameron-Pesant, Sarah; Jansen, Yorrick; Larivière, Vincent*

Research leadership in key fields for emerging and developing countries. *González-Alcaide, Gregorio; Huamaní, Charles; Park, Jinseo*

Mass Gathering as an emerging field: a co-citation analysis. *González-Alcaide, Gregorio; Llorente, Pedro; Ramos, José M.*

Research Activity Classification based on Time Series Bibliometrics. *Kawamura, Takahiro; Yamashita, Yasuhiro*

Inclusion of Gender perspective in scientific publications in Energy Efficiency. *Mauleón, Elba; De Filippo , Daniela*

Gender-based differences in German-language publications. *Mayer, Sabrina*

Scientific productivity and the impact of neurosurgery scientists in WOS and Mendeley: a gender study. *Sotudeh, Hajar; Dehdarirad, Tahereh; Pooladian, Aida*

How is the counting method for a publication or citation indicator chosen? *Gauffriau, Marianne*

The occurrence areas of the dependence problem of the h-index. *Liu, Chichen; Cai, Sanfa; Liu, Yuxian*

Multivariate bibliometric analysis of scientific production indicators: a taxonomy of countries scientific degree of centrality. *Silva, Deise D.; Grácio, Maria C. C.*

12:30-14:30 **POSTER SESSION 2**

DATA CHARACTERISATION, CLASSIFICATION, VISUALISATION AND INDICATOR DESIGN

Uncriticized citation process of the indicators like social contagion – a case study of the success rate of commercialization of the public R&D in South Korea Park. *Jinseo; Kim, Sun-Woo; Lee, June Young; Song, Tae Ho*

A comparative analysis of Western Europe and Latin America based on social and scientific indicators. *Castanha, Renata; Grácio, Maria C. C.*

Indicators of endogamy and reciprocity in PhD theses assessments. *Castelló-Cogollos, Lourdes; Aleixandre Benavent, Rafael; Castelló-Cogollos, Rafael*

Scientific Impact Indicators: a comparative study of Brazilian journals' impact factors. *Almeida, Catia C.; Grácio, Maria C. C.*

Sub-fields of Library and Information Science in Turkey: A Visualization Study. *Taşkın, Zehra; Doğan, Güleda; Al, Umut*

Content words as measure of structure in the science space. *Lamers, Wout S.*

Study on the International and Domestic Subject Areas' Distributions. *Wenjie, Wei; Junlan, Yao; Liu, Yuxian*

Characteristics of Paper Publication by Major Countries Focusing on Journals. *Fukuzawa, Naomi*

4D Specialty Approximation: Ability to Distinguish between Related Specialties. *Rons, Nadine*

Analysis of Structure of Scientific Publications at Universities Focusing on Sub-Organizations. *Murakami, Akiyoshi; Saka, Ayaka; Igami, Masatsura*


Welcome!
#sti2016vlc
www.sti2016.org


1
5
JOURNAL OF DATA AND INFORMATION SCIENCE

PLENARIES

You can watch the plenaries on youtube:

https://www.youtube.com/watch?v=uEmouz4TOds&list=PLBAoGA5erbiU9e9njUnLBBC0VhgbUHBzb

1st PLENARY SESSION

SAKIKO FUKUDA-PARR

THE NEW SCHOOL, NEW YORK, US

Sakiko Fukuda-Parr is a development economist interested in human development and the broad question of national and international policy strategies. She is currently a Professor at The New School, in the International Affairs Program where she chairs the Development Concentration. From 1995 to 2004, Sakiko was lead author and director of the UNDP Human Development Reports. Previously, she worked at the World Bank and UNDP on agriculture, aid coordination in Africa and capacity development. Recently, United Nations Secretary-General Ban Ki-moon announced the appointment of Prof Fukuda-Parr as a member of the newly established high-level panel on health technology and access to medicines.

One of her current research projects is "The Power of Numbers: A Critical Review of MDG Targets for Human Development and Human Rights (co-coordinator with Alicia Yamin, Harvard University) – a multi-author research initiative on the impact of global goal setting on international development agendas".

2nd PLENARY SESSION

14 Septembre, 16:30 – 18:00 · Room 1 · ground floor

Roundtable: Infrastructures for Inclusive and Open Science and RISIS presentation

Panelists: Éric Archambault (Science-Metrix. Montréal, Canada), Valentin Bogorov (Thomson-Reuters. Moscow, Russia), Abel Packer (Scielo, Sao Paulo. Brazil), Hebe Vessuri (IVIC. Venezuela).

Chair: Ismael Ràfols (INGENIO, CSIC-UPV. València, Spain)

The infrastructure for information on S&T has a strong influence on the patterns of communication and the visibility of science. Scientific journals and the bibliographic database shape the production, circulation and consumption of knowledge. Since the mid 20th century, science dynamics was influenced by Garfield's notion that a small "core journals" that published most of the all the research of significance – those covered by the ISI (now Web of Science) database. These core journals of 'international' scope that 'controlled' most scientific communication were mainly published in a few Western countries. The databases were often used by managers to stratify science into high quality cores (top quartile journals), second class science and 'invisible science'.

Since the 1980s, researchers in the global south and in some disciplines such SSH have increasingly voiced discontent about Garfield notion of 'core', in particular about its consequences in terms of the invisibility of 'peripheral' journals and the effects of journal stratification on knowledge production. For example, there have been worries of suppression of research on topics relevant to developing countries or marginalised populations which are published in local journals in languages other than English.

Also, the great changes in ICT in the last two decades have facilitated the pluralisation of scientific information. The appearance of new databases, such as Scielo or Redalyc that explicitly aim to fill in gaps in coverage. Moreover, the advent of open access technologies that can make 'local' journals accessible across the globe. Also new forms of science dissemination, such as blogs or twitter, or new forms of publishing (e.g. data sharing), are making scientific information more diverse. However, this succession of transformations towards more 'open science' poses major challenges to the governance of information infrastructure.

In this round table we aim to discuss, first, the diverse strategies for developing infrastructure with an open and comprehensive coverage and, second, the governance of the scientific information infrastructure in the face of new forms of communication.

First, current general databases have a limited coverage while more comprehensive databases are specific to some regions or sectors. Thus, most S&T indicators and benchmarking are based on conventional 'core' databases. Should more comprehensive databases be developed, mixing different types of science – e.g. more 'local' and more 'universal'? How should indicators of these databases be interpreted? How is open access best provided and maintained?

Second, the development of robust and publicly trusted indicators needs an open and transparent data infrastructure. What type of governance should be established to ensure public critical analysis? Which types of organisations should manage the data? Should these be distributed or centralised systems?

Previous studies of standards and infrastructure have shown that deep political implications of apparently technical choices. If we aim to make science more open, democratic and inclusive, we need to be highly reflective on how we develop these infrastructures.

3rd PLENARY SESSION

15 Septembre, 12:30 – 13:30 · Room 1 · ground floor

Global networks, internationalization and local research agendas: indicators for benchmarking or context specific?

Panelists: Jonathan Adams (Digital Science, London, UK), Rigas Arvanitis (Director of IFRIS, IRD, Paris, France), Sami Mahroum, (INSEAD Innov. and Policy Initiative, Abu Dhabi, United Arab Emirates), Mónica Salazar (InterAmerican Develoment Bank, Bogotá, Colombia).

Chair: Richard Woolley (Ingenio, CSIC-UPV, València, Spain).

It is widely accepted that 'global science' or the globalization of scientific work, collaboration and coordination has developed rapidly in the era of mass long-haul travel and has intensified with the arrival of the 'Internet age'. The ideal of a global science network through which access and contribution to science is no longer structured by zones of inclusion and exclusion is said to be within reach. In this so-called 'flat-earth' view of globalized science, physical location and local resources are secondary to international networks. Strategies for raising scientific quality are contingent on plugging into the global networks. Through these networks, countries with lower resource levels (human capital, research infrastructure, financial) are expected to access advance knowledge and techniques. This is assumed to lead to a faster rising level of competence underpinning the advancement of a science and innovation driven mode of socio-economic development.

Indicators of 'internationalization' thus become important for monitoring global connectedness as a proxy for a network model of development. Countries that map and understand their collaborations can leverage their strengths and use policy interventions to build global links in targeted areas. Indicators play an important role in highlighting opportunities and progress in connecting to key global channels. Research quality is assumed to rise in concert with internationalization indicators, lifting downstream activities and opportunities for commercial exploitation. Indicators that seek to benchmark or produce universalized measures (such as the global university rankings) are therefore regarded as relevant and seen as having positive impacts on the direction of policy development.

In contrast to this vision of global equalization, another interpretation of the globalized organization of science sees the global networks as a perpetuation of asymmetric relations of power and control over the scientific agenda. In this view, global networks mainly operate to export the research agenda of the rich and successful countries to distributed research groups in other locations. The development of a science that is not just of high quality but also of relevance to its context may be hampered by focusing on the research questions which are of interest to researchers and funding agencies in highly developed countries.

Indicator development faces other challenges according to this view that the scientific world is very far from being 'flat'. Different types of indicators might be needed in different contexts. 'Universal' measures such as global rankings may be useless, or even potentially misleading, in terms of shaping policy agendas in these contexts.

Taking these polar views, we can see that the same global network could be interpreted in two very different ways. Perhaps the challenge is to find the complementarities between these two visions. Perhaps a more reflexive politics of responsible indicator development is needed. What exactly should be the role of state administrations in this contested terrain, including those charged with capturing and presenting data for S&T information systems? This session will bring these issues of the global and the local/regional into focus and into question. It will provide an opportunity for robust debate and for challenging perspectives on the received vision of 'global science' and the indicators of internationalization that help to construct this vision.

JOHANN MOUTON

STELLENBOSCH UNIVERSITY, SOUTH AFRICA

Johann Mouton is Professor in and Director of the Centre for Research on Evaluation, Science and Technology at Stellenbosch University and the African Doctoral Academy. Johann Mouton is also the Programme Director of five post-graduate programmes in Monitoring and Evaluation Studies and Science and Technology Studies. He has authored or co-authored 10 monographs including Understanding social research (1996), The practice of social research (2002, with E. Babbie) and How to succeed in your Masters and doctoral studies (2001). He has supervised or co-supervised 70 doctoral and master's students. He received two prizes from the Academy for Science and Arts in South Africa including one for his contribution to the promotion of research methodology in South Africa. In 2012 he was elected to the Council of the Academy of Science of South Africa.

His main research interests are the philosophy and methodology of the social sciences, higher education knowledge production, sociology of science, scientometrics and science policy studies.

5th PLENARY SESSION

16 Septembre, 16:30 – 18:00 · Room 1 · ground floor

Roundtable on “Use of indicators in policy and inclusive metrics”

Panellists: Richard Deiss, (Directorate General for Research and Innovation, European Commission), Diana Hicks (Georgia Tech. Atlanta, USA), Slavo Radosevic (UCL. London, UK), Judith Sutz (President of Globelics & Univ. de la República. Montevideo, Uruguay).

Chair: Jordi Molas-Gallart (INGENIO (CSIC-UPV), Spain).

The STI conferences have long aimed to stimulate reflection on the use of indicators. Two years ago, in a plenary roundtable on “quality standards for evaluation indicators” Diana Hicks launched the idea of a “manifesto” that would lay out some basic principles on the evaluative use of indicators. This led to the Leiden Manifesto for Research Metrics, a set of “ten principles to guide research evaluation”. The Leiden Manifesto has become an influential initiative to raise awareness of the challenges posed by the use of indicators in evaluation and, therefore, to inform policy decisions. The HEFCE report The Metric Tide also recommended general principles such as robustness, humility, transparency, diversity and reflexivity regarding the responsible use of research metrics. Yet, although these principles have been well received, in many cases they do not provide solutions but state desirable goals. Agreement with the principles does not imply the capacity to implement them. How can we move from general principles to more specific advice?

This closing roundtable will discuss how to address the challenges posed by the use of indicators in policy, in particular in relation to geographical, cognitive or social areas that are not well described by current indicators.

First, we need to consider how indicators are used in the policy process. There is agreement among many evaluation practitioners that “quantitative evaluation should support qualitative, expert assessment”, as stated by the first principle of the Leiden Manifesto. Indicators and the analyses based on indicators should therefore inform but not substitute judgement. How can the principle operate in practice? Is this applicable in all circumstances? Can the application of mixed methods to evaluation help address this problem?

A second challenge relates to the adequacy of currently available indicators for assessing institutions or research against their stated missions and their specific context. The indicators community has developed sensible methods for measuring performance against some missions in certain contexts. However, some fields, such as the Humanities, or missions, such as health care, and many regions, are currently poorly covered by indicators. How can we use indicators to inform policy when they are known to be biased, for example due to the uneven topic or country coverage of databases? How should we use indicators so that local research and innovation is made visible and valued? How can we, for instance, use indicators to capture the performance of an organisation against its research missions when these are peculiar to a local context? What are the opportunities for the development and use of alternative indicators that are inclusive of currently invisible or marginalised research and innovation?

We would like to invite the panellists and the audience to share ideas and collective initiatives so that our community can contribute to a wiser, more inclusive and responsible use of S&T indicators.

ORAL PRESENTATIONS

CHAPTER 1

Data Infrastructure and Data Quality

21st International Conference on Science and Technology Indicators | València (Spain) | September 14-16, 2016 DOI:

Must Metrics Serve the Audit Society? Addressing Marketization in Open Access Publishing and Humanities Analytics[1]

Christopher Newfield*, Christopher Muellerleile**

*cnewf@english.ucsb.edu
Department of English, University of California, Santa Barbara, CA 93106 (USA)

** c.m.muellerleile@swansea.ac.uk
Department of Geography, Swansea University, Singleton Park, Swansea, SA2 8PP

ABSTRACT

The "audit society" (Power 1997) seemed to arise from its efforts to increase accountability and efficiency in public institutions. Accounting and its numerical indicators was to be a neutral tool to accomplish audits, which in turn were to increase fairness and transparency in the institutions of democratic society.

Since then, two related activities have been proceeding in parallel: the technical and institutional development of indicators, and the analysis of their institutional and sociocultural effects. Our starting point is the distance and frequent divergence of these two activities: while indicators have advanced and proliferated, their impact has been subject to largely negative critique. Most academics are habituated to ubiquitous assessment. And yet complaints about this are equally ubiquitous. Can these practices—numerical assessment and its critique—be brought into a productive relationship?

On the surface, the answer seems to be an obvious yes. Both producers and consumers of indicators release guidelines and standards designed to spread knowledge of the limits of numerical indicators and to reduce abuse (Archamabault and Gagné 2004; NICE 2013). Such guidelines invariably call for the embedding of quantitative metrics in the appropriate institutional and professional contexts. For example, the important "Leiden Manifesto" has as its first principle, "Quantitative evaluation should support qualitative, expert assessment" (Hicks et al., 2015). Similarly, the editor of *Times Higher Education*, which offers elaborate university ranking services, insists on the great value of metrics as long as consumers realize that "contextual information is vitally important" (Gill 2015). Most academics agree that valid numerical indicators can be constructed (Gingras 2014) and can be used correctly to assess research impact and productivity in the context of "informed peer review" (Wouters 2014). Core principles for the valid use of metrics are: (1) maintain the specificities and purposes of the evaluative context; (2) link quantitative to qualitative analysis; (3) include

[1] This work was supported by the Universities in the Knowledge Economy (European Commission Initial Training Network) and by the Limits of the Numerical research project (Universities of Cambridge, Chicago, and California at Santa Barbara).

professional expertise and substantive domain knowledge; (4) use data and process transparency to bind evaluators and evaluees in a shared community. And yet when institutional analysts look at the impact of numerical assessment over time, do they find these principles in action?

This paper argues that the root problem with quantitative metrics is marketization as we define it here—that although quantitative information can be used in a wide range of ways, indicators have in general allowed themselves to be caught up in the eclipsing or replacing of professional networks with marketized information systems (Mirowski 2013). To make this case, we look at two contrasting arenas: the commercialization of Open Access publishing (Muellerleile), and an emerging humanities discipline which lacks commercial potential (Newfield).

In fact, they do not. Metrics have become indissociable from "audit culture" (Shore and Wright, 2015), and the critique of audit culture is a mature field with at least two decades of work behind it. Shore and Wright (2015, 430-31) offer a summary description of its effects:

> 1. Loss of organizational trust (O'Neill 2002; Power 1994);
> 2. Elaborate and wasteful gaming strategies (House of Commons 2004; Shore and Wright 2000; Wright 2009);
> 3. A culture of compliance and large compliance costs, including the appointment of new specialists preoccupied with creating positive (mis)representations of performance (Miller 2001);
> 4. Defensive strategies and blamism that stifle innovation and focus on short-term objectives over long-term needs (Hood 2002);
> 5. Deprofessionalization, a disconnect between motivation and incentives, lower employee morale, and increased stress and anxiety (Bovbjerg 2011; Brenneis, Shore, and Wright 2005; Wright 2014);
> 6. "Tunnel vision" and performing to the measure, with a focus solely on what is counted, to the exclusion of anything else (Townley and Doyle 2007);
> 7. And the undermining of welfare and educational activities that cannot be easily measured (King and Moutsou 2010).

Since the authors of guidelines for the use of metrics (e.g. Hicks et al) would likely object to any of these seven features, not to mention their combination, why has audit practice not only survived but thrived despite this critique?

Our paper offers two cases in which we explore the marketization imperative. The systems of measuring and ranking have become objects of economic development themselves, driving innovation and the construction of new firms and markets, many of which cross over the boundaries of what was once a more autonomous university (Komljenovic and Robertson 2016). This is exacerbated by at least two things: austerity or the constant push to "do more with less," and the related managerial fascination with "big" data and "evidence driven" decision making as offering a whole new level of economization. In other words, while there is a political or ideational project at work in the process of neoliberalizing universities, there is also a material restructuring around the construction of new commodities, new markets for those commodities, and new management structures to control this new economy.

Our first case is Open Access publishing. OA has evolved through three main arguments that continue to shape the discourse today. The first is that properly functioning democratic societies are dependent upon the free circulation of knowledge (Stiglitz 1999). The argument is particularly vehement where public resources fund research, but must also pay to access the results. The second and related argument suggests that academic research and knowledge are major drivers of innovation (Stiglitz 1999, Howells et al. 2012). As such the results should be easily accessible to fuel economic competitiveness and growth. The third was a reaction to the "serials crisis" (House of Commons 2004) or the inability of university libraries to afford access to all the journals deemed necessary, coupled with what are widely seen as unjustified profits accumulated by publishers (Ciancanelli 2007). OA is to rectify access and cost problems and enable the circulation of knowledge that supports both innovation and democracy.

But both the "gold" and the "green" variants of OA have problems. In particular, they have trouble justifying the expense of peer review and the maintenance of a stable knowledge archive. They are encouraging moves toward a "publish first, filter second" mode, and then resolving questions of relevance, evidence, and overall quality posed by the lack of prior review by developing and selling bibliometric tools to filter content after relatively unreviewed publication. These tools serve two broad functions, although they are often co-constitutive. First, publishers are developing technologies that categorize, codify, and measure research and researchers. And second, publishers are using these tools to enclose, and sell meta-data about research. Through an evolution of internet media provision, what Mansell (1999) calls the "scarcity-abundance dialectic," the largest academic publishers are losing control of content, but at the same time becoming massive data aggregating corporations.

The assumption that is built into most advocacy of open access is that scientific, or just academic knowledge must be free for the public to read, if not free to put to use in any way they see fit. On the surface, this seems quite reasonable. But in order to achieve this within the current technical-economic conjuncture, the existing structures that organize academic knowledge and make it meaningful are being dismantled. In turn this is threatening to further alienate universities from the very people who open access advocates claim to have in their interest. Stated differently, capital in the form of subscription based publishers, have historically enclosed knowledge behind pay walls and copyright, but in the process they also helped to make knowledge robust and meaningful. Open Access advocates might argue that in a world of open knowledge and data, a simple Google search will solve the problem by identifying the most popular, well-connected, or most trusted research. The problem is that the for-profit publishers are ahead of this game. They are working very hard to set the rules by which Google or Mendeley or Scopus will identify the "best" academic research on any given topic. Put another way, the information structures of the Internet are not flat. They are always already filtered, curated, and uneven. Furthermore, the algorithms that control these searches are increasingly hidden from human view, or are too complicated for humans (e.g. academic researchers, academic administrators, the broader public) to understand without the aid of digital technology (Gitelman 2013, Pasquale 2015).

In spite of its potential to make knowledge more accessible, OA is being marketized in a way that will re-trigger the critique of audit culture, which details the negative institutional effects of separating research management from research work. OA can be imposed on researchers by institutional authorities, who will adapt or even embrace it. These do not reduce the intellectual or institutional costs that the critique chronicles.

Our second case is an emergent interdisciplinary U.S. field called Critical University Studies (CUS). CUS was developed and identified by humanities scholars, particularly from literary and cultural study (LCS) (Newfield 2008; Williams 2012). Its findings are of potential importance for the sociocultural life of global universities, but have no obvious market potential. The study of higher education has been marginal in U.S. education departments and schools, and has been shared among historians, sociologists, anthropologists, philosophers, accountants, management professionals, science and technology scholars, and literary critics, among others. How would an interdisciplinary terrain, rooted in informal social networks, be tracked by bibliometrics as it emerges into a new para-discipline?

This paper will report on the results of a comparison of Thomson Reuters Citation Index results to an analysis of the professional circulation of one of the key concepts developed by CUS, "cross subsidies" for extramural research conducted at U.S. universities. Conventional campus wisdom in the 2000s was that Science, Technology, Engineering and Mathematics (STEM) research generated positive revenues that subsidized money-losing fields in the arts, social sciences, and humanities. The literature on this topic was sparse and often trapped in local contexts, such as university-specific reports (UCPB 2010) or findings for membership audiences (COGR 2013). This situation began to change after 2008 through a small number of non-standard statements and writings—a quotation of a budget expert in the *New York Times* (Lieber 2009) which prompted a letter to a university president and a specialty-newspaper article (Watson 2010), and an article about budgeting in a literary journal (Newfield 2009). Many lectures were given that included references to this information (Newfield delivered 120 lectures from late 2011 through mid-2015). Over a period of 6-7 years, the conventional wisdom on cross-subsidies was reversed in arts and humanities circles, and at least challenged in STEM circles. But much of this change was social and word-of-mouth, contained in discussions and debates in department meetings and at dinner parties, and prone to resurface in publication sans attribution (Dinsman 2016). We posit that CUS represents a common pattern for emerging professional knowledge, in which such knowledge adapts to local contexts and evolves as it migrates through affinity networks, while tending to shy away from the high-profile venues most concerned with their reputation and markets. As a result, emerging knowledge evades marketization, and can be undermeasured. Its circuits function more like an artistic avant-garde or a musical subculture (Hebdige 1979) than like a citational network.

Through empirical descriptions of the circulation of CUS knowledge, we will suggest that core axioms of Garfield bibliometrics do not function normally in a non-marketized system of professional knowledge. Here we do *not* find that citation frequency is a reliable index of cognitive impact (De Bellis 2009); that citation indexes measure the intellectual impact and productivity of individuals and units over time, and thus can be used for quality assurance and other management functions (Hirsch 2005); that the standard concentration of references to a small number of scientists reflects the social actuality of knowledge generation and influence; that influence follows a power law function, or follow Pareto rather than Gaussian distribution (Lotka, 1926; Bradford 1934; Zipf, 1936; De Bellis, 2009). Although the study of literature and the arts has sometimes adopted a belief in the concentration of genius by focusing on canonical masterworks, this has decreasingly been the case in recent decades, as the humanities fields have studied the circulation of aesthetic and cultural intelligence throughout entire social systems, particularly across cultural differences and around the alleged peripheries.

Our comparison of these two arenas, OA and CUS, will allow us to conclude with suggestions for how bibliometrics might track distributions of knowledge that are neither Gaussian nor Paretian, but informal, subterranean, and democratic. This may be an opportunity for quantitative assessment to part company with audit culture.

REFERENCES

Archamabault, Éric and Étienne Vignola Gagné. 2004. *The Use of Bibliometrics in the Social Sciences and Humanities*. Montréal: Social Sciences and Humanities Research Council of Canada.

Bovbjerg, Kirsten-Marie, ed. 2011. "Motivation og mismod: effektivisering og stress på offentlige arbejdspladser." Aarhus, Denmark: Aarhus University Press.

Bradford, Samuel C. 1934. "Sources of Information on Specific Subjects," *Engineering* 137(3550): 85-86.

Brenneis, Don, Cris Shore, and Susan Wright. 2005. "Getting the measure of academia: universities and the politics of accountability." *Anthropology in Action* 12(1):1–10.

Ciancanelli, P. 2007. "(Re)producing universities: Knowledge dissemination, market power and the global knowledge commons." In Epstein, D., Boden, R., Deem, R., Rizvi, F., and Wright, S. (eds.) *World Yearbook of Education 2008*, New York: Routledge

COGR (Council on Governmental Relations). 2014. "Finances of Research Universities." May. http://cogr.edu/COGR/files/ccLibraryFiles/Filename/000000000267/Finances%20of%20Research%20Universities_June%202014.pdf

Cronin and Cassidry R. Sugimoto, eds. 2014. *Beyond Bibliometrics: Harnessing Multidimensional Indicators of Scholarly Impact.* Cambridge, Mass.: MIT Press.

De Bellis, Nicola. 2009. *Bibliometrics and Citation Analysis: From the Science Citation Index to Cybermetrics*. Lanham, Md.: The Scarecrow Press.

Dinsman, Melissa. 2016. "The Digital in the Humanities: An Interview with Alexander..." *The Los Angeles Review of Books*, March 27. https://lareviewofbooks.org/interview/the-digital-in-the-humanities-an-interview-with-alexander-galloway/.

Gill, John. 2015. "Taking Appropriate Measures," *Times Higher Education.* 9 July 2015: 5.

Gingras, Yves. 2014. "Criteria for Evaluating Indicators." In Cronin and Sugimoto, eds. 109-25.

Gitelman, L. ed. 2013. *"Raw Data" is an Oxymoron*, Cambridge, MA: MIT Press

Hebdige, Dick. 1979. *Subcultures: The Meaning of Style.* London: Routledge.

Hicks, D., Wouters, P., Waltman, L., de Rijcke, S., & Rafols, I. (2015). The Leiden Manifesto for Research Metrics. *Nature*, *520*, 429–431.

Hood, C. 2002. "The risk game and the blame game." *Government and Opposition* 37(1):15–37.

House of Commons. 2004. "Scientific Publications: Free for all?" House of Commons Science and Technology Committee, Tenth Report of Session 2003-2004. London: Her Majesty's Stationary Office:http://www.publications.parliament.uk/pa/cm200304/cmselect/cmsctech/399/399.pdf

Howells, J., Ramlogan, R., and Cheng, S.L. 2012. "Innovation and university collaboration: paradox and complexity within the knowledge economy." *Cambridge Journal of Economics*, vol. 36: 703-721.

King, Lucy, and Christina Moutsou, eds. 2010. *Rethinking audit cultures.* London: PCCS Books.

Komljenovic, Janja & Susan L. Robertson (2016): The dynamics of 'market-making' in higher education, *Journal of Education Policy*

Lotka, Alfred J. 1926. "Statistics: The Frequency Distribution of Scientific Productivity." *Journal of the Washington Academy of Sciences* 16 (12): 317-25.

Lieber, Ron. 2009. "Why College Costs So Much." *The New York Times*, September 4. http://www.nytimes.com/2009/09/05/your-money/paying-for-college/05money.html.

Miller, Peter. 2001. "Governing by numbers: why calculative practices matter." *Social research* 68(2):379–396.

Mansell, R. 1999. "New media competition and access: the scarcity-abundance dialectic", *New Media and Society*, vol. 1, iss. 2: 155-182.

Mirowski, Philip. 2013. *Never Let a Serious Crisis Go to Waste: How Neoliberalism Survived the Financial Meltdown*. London ; New York: Verso.

Newfield, Christopher. 2008. *Unmaking the Public University: The Forty-Year Assault on the Middle Class*. Reprint. Harvard University Press.

Newfield, Christopher. 2009. "Ending the Budget Wars: Funding the Humanities during a Crisis in Higher Education." Modern Languages Association: *Profession*.

NICE. 2013. *Guide to the methods of technology appraisal*. London: National Institute of Health and Clinical Excellence http://publications.nice.org.uk/pmg9

O'Neill, Onora. 2002. "A question of trust: the BBC Reith lectures of 2002." Cambridge: Cambridge University Press.

Power, Michael. 1994. *The audit explosion.* London: Demos. http://www.demos.co.uk/files/theauditexplosion.pdf

Power, Michael. 1997. 1997. The audit society: rituals of verification. Oxford: Oxford University Press.

Shore, Cris and Susan Wright. 2015. "Audit Culture Revisited: Rankings, Ratings, and the Reassembling of Society. *Current Anthropology* 56: 3, June.

Stiglitz, J. 1999. "Knowledge as a Public Good." In Kaul, I, Gruenberg, I. and Stern, M.A. *Global Public Goods: International Cooperation in the 21st Century*. Oxford: Oxford University Press

Townley, Barbara, and Rosemary Doyle. 2007. "Performance measurement." In *Blackwell Encyclopedia of Sociology*. George Ritzer, ed. Blackwell Reference Online. http://www.blackwellreference.com/subscriber/tocnode.html?idpg9781405124331_chunk_g978140512433122_ss1-17

UCPB (University Committee on Planning and Budget). 2010. "Choices Report." Oakland: University of California Office of the President. http://senate.universityofcalifornia.edu/reports/ucpb.choices.pdf

Watson, Robert N. 2010. "The Humanities Really Do Produce a Profit." *The Chronicle of Higher Education*, March 21. http://chronicle.com/article/The-Humanities-Really-Do/64740/.

Williams, Jeffrey J. 2012. "Deconstructing Academe." *The Chronicle of Higher Education*, February 19. http://chronicle.com/article/An-Emerging-Field-Deconstructs/130791/.

Wright, Susan. 2014. "Knowledge that counts: points systems and the governance of Danish universities." In *Under new public management: institutional ethnographies of changing front-line work*. Dorothy Smith and Alison Griffith, eds. Toronto: University of Toronto Press.

Wright, Susan. 2005. "Processes of social transformation: an anthropology of English higher education policy." In Pædagogisk Antropologi—et Fag I Tilblivelse. J. Krejsler, N. Kryger, and J. Milner, eds. Pp. 185–218. København: Danmarks Pædagogiske Universitets Forlag.

Wouters, Paul. 2014. "The Citation: From Culture to Infrastructure." In Cronin and Sugimoto, eds. 47-66

Zipf, George K. 1935. *The Psycho-Biology of Language: An Introduction to Dynamic Philology*. London: Routledge, 1936.

On the Peripheries of Scholarly Infrastructure: A look at the Journals Using Open Journal Systems1

Juan Pablo Alperin*, Kevin Stranack** and Alex Garnett**

*juan@alperin.ca
Canadian Institute for Studies in Publishing, Simon Fraser University, 515 West Hastings Street, Vancouver, 6B 5K3 (Canada)
Public Knowledge Project, Simon Fraser University, 8888 University Drive, Burnaby, V5A 1S6 (Canada)

** kstranac@sfu.ca; garnett@sfu.ca
Public Knowledge Project, Simon Fraser University, 8888 University Drive, Burnaby, V5A 1S6 (Canada)

ABSTRACT
Although there have been calls for scholarly infrastructure to be inclusive, new layers of infrastructure are built without a clear understanding of the breadth of scholarly journals that lie on the peripheries of the existing infrastructure. In the hopes that future infrastructure can take a wider range of journals into account, this paper presents the results of an effort to track the number, location, and rate of publication of journals using Open Journal Systems, an open source manuscript management and publication system built by the Public Knowledge Project. The method employed, which involves a combination of scanning weblogs, scraping webpages, and harvesting metadata, has yielded an estimated 9,828 journals that have collectively published 2,565,300 articles since 1990. These journals are distributed across 136 countries on 6 continents, and, in 2015, around a fifth of the OJS journals were published in low or low-middle income countries, and over a third in upper-middle income countries, suggesting that the majority of OJS journals are on the on the “periphery” of today's global scholarly infrastructure. As infrastructure and services continue to be developed, this paper argues, it is necessary to look to such journal so that the infrastructure that is built can be developed in a way that is truly inclusive of the global nature of scholarship.

BACKGROUND
The Public Knowledge Project (PKP) is a research and development initiative of Simon Fraser University and Stanford University, with a focus on understanding and building enhanced modes of scholarly communication that facilitate open access, high quality publishing, and local capacity building and participation.

One of the most significant contributions from PKP has been the development of its free, open source Open Journal Systems (OJS) software.[i] OJS is a professional journal publishing platform that is easy to download, install, and operate with minimal server requirements. OJS allows for online submissions, peer review, copyediting and layout, and publishing. It also provides connections to indexing (e.g., PubMed, DOAJ), digital identifiers (e.g., CrossRef, ORCID), and preservation services (e.g., LOCKSS, Portico).
Because OJS is free and is designed to maximize efficiencies for publishing activities, publishers have been able to launch journals that would otherwise have been financially untenable. Although some existing publishers have made use of the system, the majority of

In the interest of full disclosure, it should be noted that the authors of this work are all affiliated with the Public Knowledge Project, the creators of Open Journal Systems.

OJS users are new to publishing and are based in academia, either through library publishing programs or by individual "scholar-publishers," determined to build communities of interest around their research areas (Edgar & Willinsky, 2010). Most are small-scale, often operating with in-kind contributions and minimal budgets (Edgar & Willinsky, 2010). This type of use suggests an increased participation in scholarship, both in terms of who publishes and who read.

In a sense, OJS can be already be said to form part of the scholarly infrastructure. It provides a layer over which scholarly activity takes place, and it brings the use of standards and best practices to those activities. It acts as one layer of infrastructure especially for those journals who would are not formally part of the scholarly infrastructure

METHODS

We have sought to use automated methods for identifying OJS journals, and subsequently use a combination of web scraping[ii] and the journal's OAI PMH endpoint[iii] to collect publicly available information about the journal, including article metadata. Because this process is entirely automated, it can be continuously run, yielding a dataset of journals and articles data that will eventually include longitudinal data. These data are then processed to identify the number of journals, articles, and their geographic location.

In a latter phase, yet to be completed (and thus not presented here) these data will be complemented with an online survey, loosely based on a similar study of OJS journals done by Edgar & Willinsky (2010), and targeted at the email addresses collected through the first phase data collection.

The first challenge in studying these journals is to learn of their existence. Because the OJS software is open source, each journal or publisher can install the software on their own server, without ever needing to register with PKP or elsewhere. Only once a journal's web address (URL) has been identified is it possible to learn how many articles it publishes in any given year, its geographic location, and its identifying information and other metadata. To collect this data, we have devised a necessarily complex method of extracting and processing the information (Figure 1).

Figure 1. Steps to collect and process OJS journal information

1. Process the PKP website logs for referrer URLs that 'look like' OJS journals
2. Attempt to contact the OAI PMH URL corresponding to the journal URL (following known OJS URL patterns) to verify if it is an OJS journal
3. Save the repository identifier, I.P. address, OJS version number
4. Identify all the journals for this installation using the OAI verb "ListSets"
5. Save the journal name, and journal contact email address from the OAI response for later use
6. Add known OJS OAI URLs to an instance of the PKP Harvester
7. Look up the journal's country
8. Collect the article metadata for every journal using OAI PMH
9. Process the article data to identify number of articles published per year, the country of origin of the journal, etc.

Note: Code for step 1 can be found at: https://github.com/pkp/ojsstats/blob/master/pkp-log-parser.php; Code for steps 2-5 can be found at: https://github.com/pkp/ojsstats/blob/master/checkOJS.py; Code for step 6 can be found at: https://github.com/pkp/ojsstats/blob/master/harvesterInsert.py; Code for step 7 can be found at: https://github.com/pkp/ojsstats/blob/master/countryLookup.py; Step 8 is handled by the PKP Harveser, which can be found at: https://github.com/pkp/harvester; Code for step 9 can be found at: https://github.com/pkp/ojsstats/blob/master/statscrunch.py. The process is coordinated by a "cron" job, which can be found at: https://github.com/pkp/ojsstats/blob/master/cron.sh. Detailed description of the each step can be found in the README: http://github.com/pkp/ojsstats.

There are several known ways in which the above system can fail to correctly identify and query a valid OJS journal. It is possible, for example, that a journal removes the links to the PKP website from their installation altogether, rendering it invisible to our logs; similarly, it is possible that journals have modified the code to change the way URLs are structured, making them difficult to correctly identify; it is also possible that journals have disabled their OAI endpoint, thereby invalidating the way in which we verify an OJS installation and subsequently collect data; or, a journal may simply be unavailable online at the time when we attempt to crawl it, leading our system to believe that the site no longer exists. Any of these circumstances, and possibly others, would result in us undercounting the number of OJS journals.

While we realize there are simpler ways of collecting this data about OJS journals (i.e., the data could be pushed from the OJS installation to a centralized system), it has always been PKP's approach to give control of the data to the journals, and to ask for nothing in return for using the software. In recent years, however, PKP has recognized the need to better understand its journal community, and, as a result, has included a "phone home" feature that provides PKP with a minimal set of publicly available data (with an opt-out option). This feature has only been available since 2015, so it was still necessary to develop the methods outlined above to produce the results that follow.

RESULTS

We identified 6,271 installations of OJS with some content, spread across 136 countries on 6 continents. These installations collectively host 9,828 "journals" that meet our arbitrary threshold of at least 18 articles published in the previous two years.[iv] At the time of writing, there were 7,491 journals that met this threshold for 2015, and 9,315 that met it for 2014 (some journals appear to add content with a delay in publication resulting in the appearance of a drop-off in numbers in 2015). Of these, the top 3 countries, by number of journals published using OJS with recent content, are 1,426 in Brazil, 1,075 in Indonesia, and 912 in the US (Figure 2). No other country had more than 500 journals that met the criteria in 2015. Interestingly, Latin America (led by Brazil) publishes approximately one third of all OJS journals.

Naturally, the corresponding country income level accompanies the geographic distribution. Approximately 42% of the journals are published by high-income countries, 38% by upper-middle income countries, and the final 20% by low or low-middle income countries (Table 2).[v]

These journals have published 2,565,300 articles since 1990. As might be expected, more articles are published using the software in more recent years, with over 300,000 articles published in 2013 and 2014, and over 250,000 in 2012 and 2015 (Figure 3). This is due primarily to an increase in the number of journals that have gone (and stayed) online over time (Figure 4). Given that the software was originally released in 2001, it seems that journals come online, add some amount of back-content (archives), and then continue to publish. In recent years (since 2010), OJS journals that meet the 18-article threshold have published 29.8 articles (up from 25.6 in the 15 years prior) (Figure 5).

Figure 2. Location of OJS journals in 2015

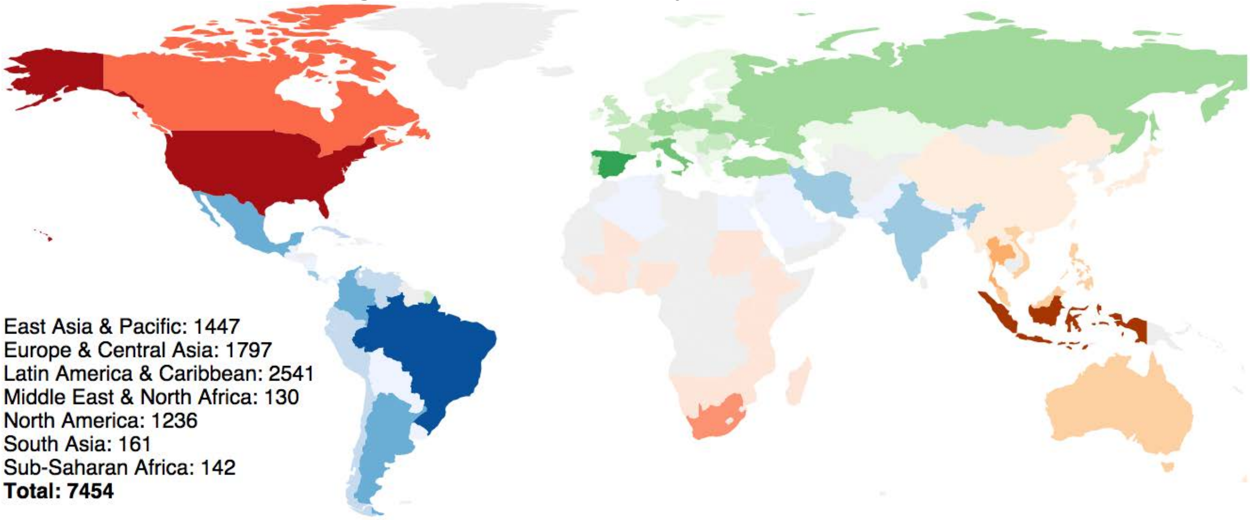


Table 2. Country Income-Level of OJS Journals for 2015

Income Category	Number and percent of journals (2015)	Number and percent of articles (2015)	Average Number of Articles per journal (2015)
Low	53 (0.7%)	1,225 (0.4%)	23.1
Lower-middle income	1,525 (19.7%)	68,571 (23.8%)	46.6
Upper middle income	2,832 (37.8%)	94,945 (33.0%)	33.5
High income	3,129 (41.8%)	123,211 (42.8%)	39.4

Figure 3. Number of Articles Published in OJS Journals By Year

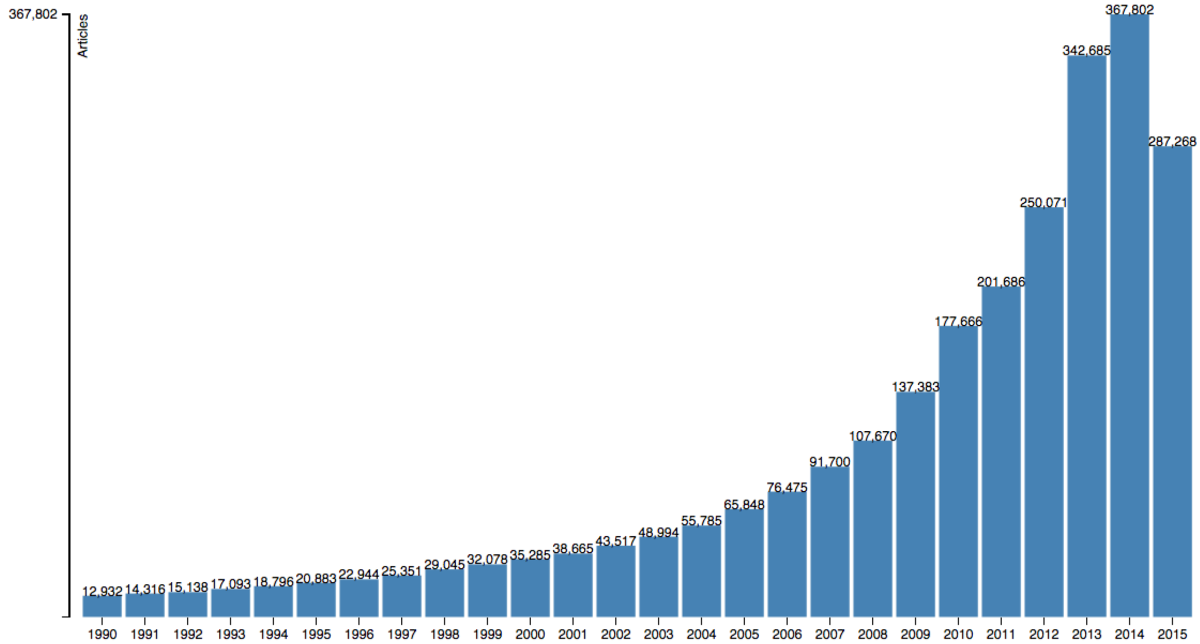


Figure 4. Number of Journals Using OJS by Year

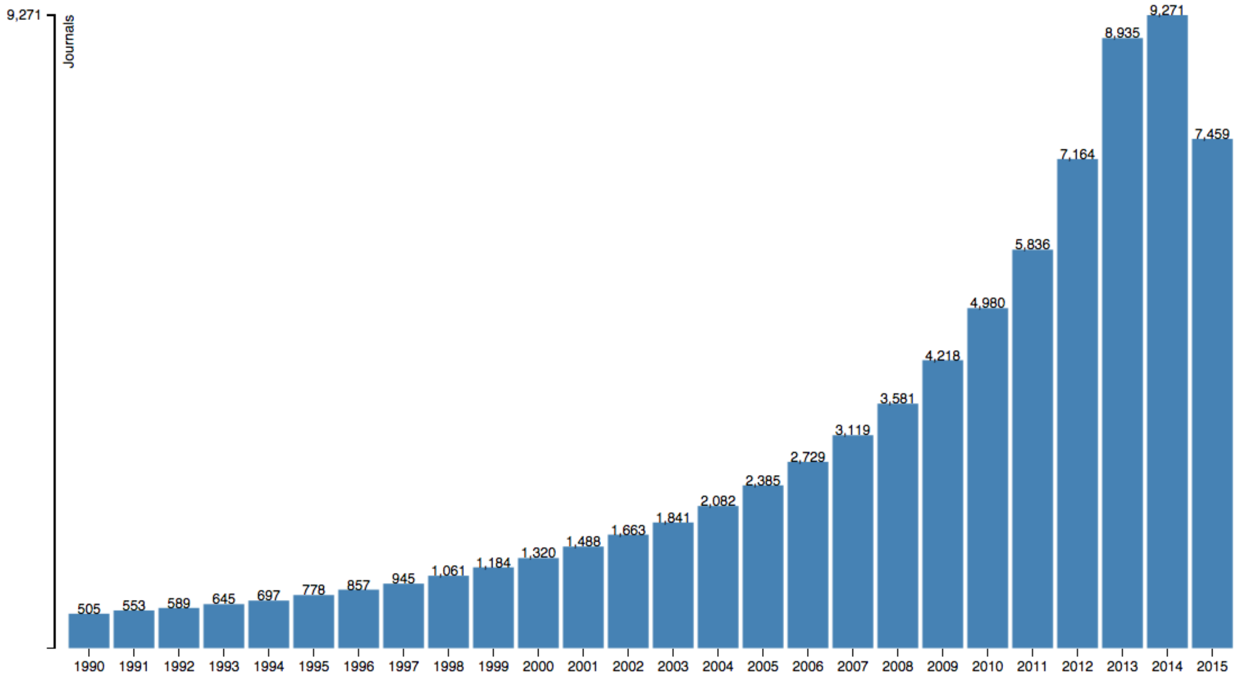
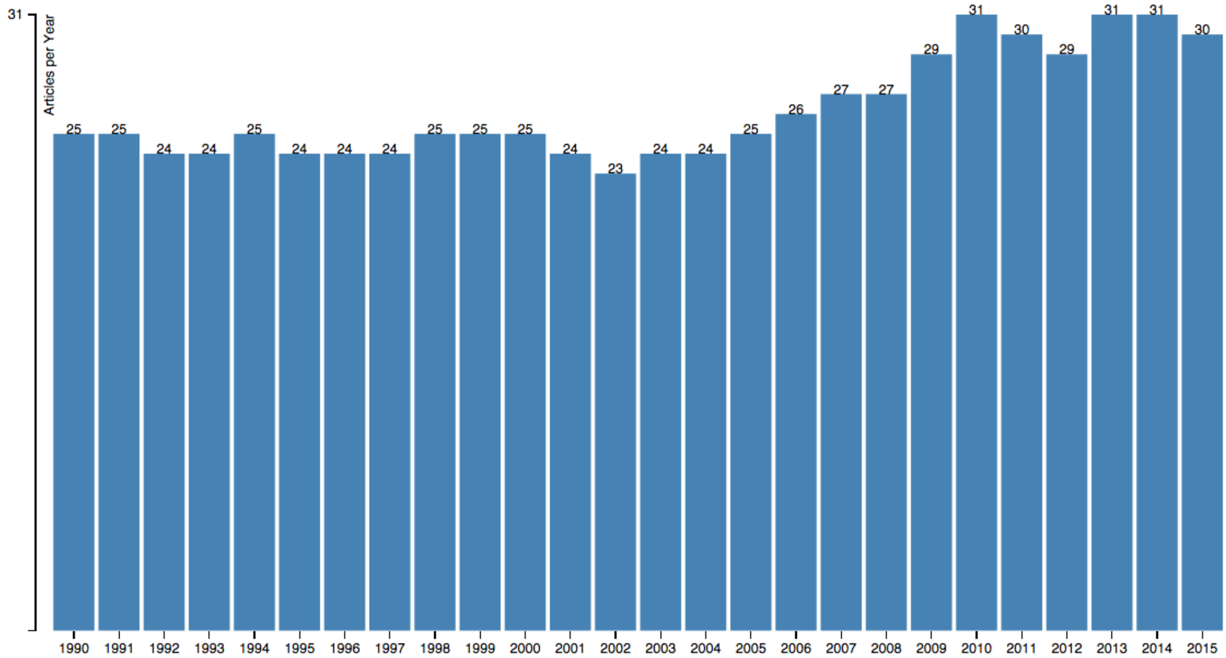


Figure 5. Average Number of Articles Published per Year in OJS Journals



DISCUSSION AND FUTURE WORK

Because of the imperfect criteria and roundabout method for collecting the data, these numbers should be treated as estimates. However, although we are surely counting some instances that are not really journals as have been traditionally understood, we are also certain that we have missed some instances altogether, and are reasonably comfortable that they provide a sufficiently accurate estimate to provide a better understanding of the type of journals that are in need of access to the scholarly infrastructure. Future studies and improved methods, including the better automated data collection, will allow us to assess these estimates.

The geographic extent and sheer number of journals and articles—most of which are not likely to be found in international citation databases—should serve to open our eyes to the need to think more broadly about who is actively trying to communicate scholarship.

The extent to which these journals are currently outside the scholarly infrastructure cannot be overstated. Not only are these journals excluded from major citation databases (there is no need to test their inclusion, as the sheer volume of journals from outside North America and Europe dwarf the total number of journals from these regions in Web of Science and Scopus), they are also lacking some of the basic elements of the scholarly infrastructure, such as DOIs. As of November 2015, there were only 837 CrossRef members that used OJS (CrossRef, personal communication). These members collectively minted 701,622 DOIs (out of over 2.5M articles).[vi] Even under the generous assumption that each DOI minted corresponds to an article (i.e., not supplementary files, data, etc.), this amounts to less than a third of the total articles published with OJS journals.

The geographic and economic distribution of access to elements of the scholarly infrastructure, such as DOIs, remains to be studied.[vii] It is nonetheless striking (although likely coincidental) that the number of DOIs minted by CrossRef members using OJS in 2015

(127,026) coincides with the number of articles published by OJS journals from high-income countries that year (123,211).[viii] While PKP works towards improving the adoption of DOIs for these journals and CrossRef works to revamp its small-publisher tools, the reality is such that the scholarly metrics and indicators that can be calculated today do not consider these journals on the peripheries of the scholarly communication infrastructure.

As can be seen from the complexity of the data collection described above, deriving metrics from these small, decentralized publishers is a challenge. Attempting to extend our efforts beyond the journals published with OJS would make the task nearly impossible. In this way, OJS is itself a piece of the infrastructure, one that allows us to account for over 2.5M articles and gives us the opportunity to learn more about them.

If these journals were part of the larger metrics infrastructure and systems, it would fundamentally change the peripheral nature of the scholars and scholarly work contained therein. As these figures highlight, there are literally millions of articles published from this periphery. The over 20% of OJS journals in low and low-middle income countries (corresponding to nearly 25% of all OJS-published articles) need to be included if the indicators and metrics are to be truly representative of the scholarship that is out there.

Being inclusive in the indicators and metrics infrastructure is necessary if we wish to use these indicators to understand how scholarship is carried out around the world. Being inclusive, however, is not as simple as putting out an open invitation to participate. It is also necessary to lower the barriers to access by making sure the infrastructure is suitable and adapted to the needs of everyone. To do that effectively, we need to know who we are trying to include, so that they can be consulted and can participate from the beginning as the infrastructure and metrics are designed and built. The research described here is a first look at many of the thousands of journals who are currently not being included.

This work is very much ongoing. In the coming months, we will conduct a survey of these journals to learn about the practices, challenges, funding, and other information that will be directly relevant to those seeking to build truly inclusive scholarly infrastructure. While this first phrase has given a general sense of the scale and extent of journals, it is only the beginning of a longer process of trying to understand the global research landscape.

REFERENCES

Alperin, J. P. (2013, October 11). Are ALMs/altmetrics propagating global inequality? Retrieved from https://speakerdeck.com/jalperin/altmetrics-propagating-global-inequality

Alperin, J.P. (2014a). Altmetrics could enable scholarship from developing countries to receive due recognition. *LSE Impact Blog.* Retrieved from: http://blogs.lse.ac.uk/impactofsocialsciences/2014/03/10/altmetrics-for-developing-regions/

Alperin, J. P. (2014b). South America: Citation databases omit local journals. *Nature*, *511*(7508), 155–155. http://doi.org/10.1038/511155c

Bilder, G., Lin, J., & Neylon, C. (2015). Principles for Open Scholarly Infrastructures-v1. http://doi.org/10.6084/m9.figshare.1314859.v1

Cetto, A. M., & Alonso-Gamboa, J. O. (1998). Scientific periodicals in Latin America and the Caribbean: a global perspective. *Interciencia*, *23*(525), 84–93. Retrieved from http://www.interciencia.org/v23_02/cetto.pdf

Collazo-Reyes, F., Luna-Morales, M. E., Russell, J. M., & Pérez-Angón, M. (2008). Publication and citation patterns of Latin American & Caribbean journals in the SCI and SSCI from 1995 to 2004. *Scientometrics*, *75*(1), 145–161. http://doi.org/10.1007/s11192-007-1841-6

Edgar, B. D., & Willinsky, J. (2010). A Survey of the Scholarly Journals Using Open Journal Systems. *Scholarly and Research Communication*, *1*(2), 1–22.

Haustein, S., Sugimoto, C., & Larivière, V. (2015). Guest editorial: social media in scholarly communication. *Aslib Journal of Information Management*, *67*(3). http://doi.org/10.1108/AJIM-03-2015-0047

Russell, J. M., Madera Jaramillo, M. J., & Shirley, A. (2009). El análisis de redes en el estudio de la colaboración científica. *Redes: Revista Hispana Para El Análisis de Redes Sociales*, *17*(2), pp. 39–47. Retrieved from http://dialnet.unirioja.es/servlet/articulo?codigo=3165613&orden=243060&info=link

Terra Figari, L. I. (2008). Diseminación del Conocimiento Académico en América Latina. In *Antropología Social Y Cultural En Uruguay 2007*. Montevideo, Uruguay: UNESCO. Retrieved from http://www.unesco.org.uy/shs/fileadmin/templates/shs/archivos/anuario2007/articulo_15.pdf

[i] http://pkp.sfu.ca/ojs

[ii] Using a script to make web requests, like a browser would, to access content from the journal pages.

[iii] OAI-PMH (Open Archives Initiative Protocol for Metadata Harvesting) is a standard used by many journal and repository systems to expose metadata in a machine-readable format.

[iv] This 18 articles in the previous two years is based on Canada's Social Science Research Council's Aid to Scholarly Journals' guideline that stipulates 12 peer reviewed articles in the previous two years (see http://www.sshrc-crsh.gc.ca/funding-financement/programs-programmes/scholarly_journals-revues_savantes-eng.aspx#a5 for details). We assume, therefore, that two thirds of articles are peer reviewed to arrive at an 18 article threshold). This minimum was set in an effort to filter out test installations, or installations that have just gone online and have not fully gotten underway. Since we are not verifying any other criteria, the word "journal" is therefore used loosely.

[v] The income levels are drawn from the World Bank classifications, with both high-income categories (OECD and non-OECD combined) and the low and lower-middle income categories combined.

[vi] Note that a member could correspond to more than one journal.

[vii] Ironically, the most suitable way of cross-checking the OJS database and others would be to use DOIs.

[viii] Even if this were not a coincidence, and the majority DOIs were minted by journals from high-income countries, it should be noted that DOIs minted in 2015 do not necessarily correspond to content published in 2015, so it is unlikely there would be a one-to-one correspondence.

Why researchers publish in journals not indexed in mainstream databases: training, bridging and gap-filling[1]

Diego Chavarro*, Puay Tang* and Ismael Rafols*, **

* diego.chavarro@sussex.ac.uk, p.tang@sussex.ac.uk
SPRU (Science Policy Research Unit), University of Sussex, Brighton, BN1 9SL, England

** i.rafols@ingenio.upv.es
Ingenio (CSIC-UPV), Universitat Politècnica de València, València, 46022, Spain
Observatoire des Sciences et Techniques (OST-HCERES), Paris, France

ABSTRACT
Although journals indexed in mainstream Journal Indexing Systems (JIS), i.e. Web of Science (WoS) and Scopus, have more visibility, reputation and are more valued in evaluation, in developing countries researchers keep publishing in journals not indexed in mainstream databases, but indexed instead in alternative JIS such as Scielo or RedALyC. The conventional explanation to this behaviour is that developing countries' research often does not have sufficient quality so as to be published in journals indexed by WoS or Scopus. We conducted 30 interviews to researchers in Colombia working in chemistry, agricultural sciences, and business and management asking the reasons for publishing in different types of journals, in particular those indexed by mainstream and alternative JIS. The answers provided by the researchers were classified into three reasons. The first reason is that journals in alternative JIS offer a space for *training* in publishing, both as an introduction to academic publishing and as a step-stone towards publishing in WoS. The second reason is that journals indexed by Scielo and RedALyC have a *knowledge-bridging* function, providing a link between articles covered by mainstream JIS and articles of regional communities with limited access to WoS or Scopus journals. The third reason is that alternative JIS journals have a *knowledge-gap filling* function, allowing the publication of topics that are not well covered in WoS-indexed journals, such as locally relevant agricultural products or regional history. We conclude that scientometric indicators based on mainstream JIS underrepresent the contribution of research from developing countries – as they do not value these training, knowledge-bridging and gap-filling functions. We discuss these findings in the light on universalistic versus particularistic conceptualisations of science.

INTRODUCTION
Why do researchers in developing countries such as Colombia keep publishing in journals not indexed by WoS or Scopus, given their low visibility, low reputation and that they are lowly valued in evaluation exercises? The explanation implicitly held by many evaluation or scientometric experts is that if they could, all researchers would publish in journals indexed in WoS, which are those with the highest quality. This belief follows from the ideas on research quality of Eugene Garfield, founder of WoS. He argued that "the significant scientific literature appears in a small core of journals" (Garfield, 1996). According to him, this core

[1] This work is based on insights from one chapter of Diego Chavarro's thesis in SPRU (University of Sussex, submitted in 2016) supported by a fellowship from Colciencias, Colombia. PT and IR provided advice as his supervisors.

was composed of around 150 journals that "account for half of what is cited and a quarter of what is published in WoS" (Garfield, 1996). The concept of core journals has been used to determine and justify the coverage of WoS. Basically, the aim of WoS is to select a portion of scientific journals characterised by their high scientific impact and their compliance with editorial standards. These characteristics can be considered as universalistic: they could be achieved by any journal regardless of its language, discipline, or country of publication. Within this rationale, the journal coverage of WoS is considered objective and the exclusion of journals is seen as justified.

If this is the case, why should researchers publish outside WoS or Scopus? The development of alternative Journal Indexing Systems (JIS) such as Scielo or RedALyC suggests that non-mainstream journals fulfil functions that are valued by researchers and policy-makers in regional contexts such as the Ibero-American. An analysis of WoS' and Scopus' coverage shows that their coverage is particularistic, meaning that geographic, linguistic, and disciplinary biases have an important impact on journal selection decisions (Chavarro, 2016, unpublished). As a result one can expect that publication in journals indexed in alternative JIS have various valuable functions beyond or rather than publishing "low quality" research (i.e. research that is not perceived as making a significant contribution to knowledge by peers in 'global' scientific communities).

What are the functions performed in the scientific system of a developing country provided by publications in journals indexed by alternative JIS? In this article we investigate these functions by examining the reasons reported in thirty interviews by researchers in Colombia for publishing in diverse journals.

Colombia as a case study

We use Colombia to examine the publishing practice of researchers in developing countries with a growing number of publications in mainstream and alternative JIS. Colombia is classified as an upper-middle income country by the OECD and usually also classified as an S&T developing country (Ordóñez-Matamoros et al, 2010). It is an important producer of scholarly journals in Ibero-America. It can be compared on its production of journals to Brazil, Mexico, Chile, Peru, Venezuela, and Cuba. As in other Ibero-American countries, most of its scholarly publishing houses are higher education institutions. However, few journals produced in Colombia are covered by WoS.

Additionally, a good number of these journals are from the social sciences and are published in Spanish. This means that Colombia has multiple disadvantages in terms of coverage by WoS. At the same time, scientists working for Colombian organisations have increased their production in journals indexed by WoS, which is a trend in Ibero-America (Lemarchand, 2012). This shows two phenomena happening in parallel: the first is the increasing production of journals indexed by alternative JIS; the second is the growing number of papers in WoS-indexed journals by Ibero-American researchers. Ibero-American researchers create these phenomena by their decisions on where to publish – WoS, alternative JIS or both. This makes them an essential source of information on why alternative JIS develop. Their position as researchers in a peripheral country to WoS makes this case valuable to understand the development of alternative JIS.

METHODOLOGY

Data source and sample

30 interviews were conducted from May to September 2013 in Colombia. The sample of researchers was taken from three different disciplines, namely chemistry, agricultural sciences, and business and management. The main reasons to choose these disciplines are extent of coverage and context of application. In terms of coverage, chemistry is generally well covered by WoS while the other two are not. This could imply a lesser need for alternative JIS in chemistry as compared to the other two disciplines.

The researchers in the sample have a variety of backgrounds that are shown on table 1:

Table 1. Distribution of researchers interviewed

Sector	Private university	19
	Public university	11
Experience	Senior	17
	Junior	13
Gender	Women	9
	Men	21
Nationality	Colombian	26
	Other	4

Importantly, these researchers exhibit different publication patterns in journals covered by WoS, Scopus, Scielo, and RedALyC. They were identified using CvLAC. This is a Curriculum Vitae database managed by Colciencias, the main public funding agency for science in Colombia. The criteria to select researchers were based on (1) those participating in a research group endorsed by a Colombian organisation certified by Colciencias[2], (2) having a PhD, and (3) having an individual production of at least three papers in the last ten years. In actuality most of the interviewees have five or more papers. We contacted 60 researchers in total, and conducted 30 formal interviews with them - ten for each discipline.

Interview protocol

The interview program was intended to answer the research question: *why do researchers publish in journals indexed by alternative JIS?* The interviews followed a semi-structured, open-ended questionnaire. A final questionnaire that we grouped into five main topics:

1. Reasons to publish research.
2. Explanation of the publication patterns of researchers in terms of JIS.
3. Use of Scielo, RedALyC, WoS, and Scopus in research.
4. The “value” of Scielo, RedALyC, WoS, and Scopus for their publications.
5. The future of JIS, recommendations, comments.

Twenty-eight of the interviews were recorded. We used the method known as thematic analysis, which consists of taking notes while interviewing, then journalizing the notes as soon as the interview is finished, listening to the audio files, identifying categories, and validating the categories found through a second review (Braun and Clarke, 2006). The

[2] In Colombia, in order to be recognised as a research group by Colciencias, the supporting organisation has to confirm it formally. This is known as endorsement.

responses were complemented by secondary data sources. Specifically, the CVs of the researchers in the sample, the examination of Scielo, RedALyC, WoS, and Scopus, and the analysis of specific papers mentioned by them.

RESULTS

The analysis of the interviews suggests three sets of reasons for publishing in journals not indexed in the mainstream databases. Details with quotes of the reasons will be provided in full article, but for lack of space, here we provide the distillation of the insights in terms of three motivations or reasons. We refer to them as *training, knowledge-gap filling,* and *knowledge-bridging.*

Training

Journals in alternative JIS-indexed are perceived as "transit stations" towards WoS-indexed journals or training arenas for initiation into publication in WoS-indexed journals. There are two bases for this argument.

a) **Journals in alternative JIS are used as training for researchers to publish in WoS-indexed journals**: The experience gained by publishing in alternative JIS-indexed journals increases the skills of researchers to publish in WoS-indexed journals. The papers they publish in alternative JIS-indexed journals incorporate this feedback, which contributes to improving the robustness of other papers that will be submitted to WoS-indexed journals in English.

b) **These journals are also used to introduce PhD students to academic publishing in their own language**: researchers encourage PhD students to look for literature and to publish papers in alternative JIS-indexed journals as part of their formation as academics. This is different from item (a) above in that the ultimate aim is not to publish in WoS-indexed journals, but to initiate new researchers into publishing. Doctoral students also get acquainted with the peer review system, regardless of their success in publishing or their future publication patterns.

In this sense, Scielo and RedALyC are seen as a means, whereas WoS is seen as the goal. This perception comes from the idea that there is a sequential publishing pattern in a researcher's career: from non-indexed through alternative JIS to mainstream JIS-indexed journals. This is a universalistic understanding of stratification of research quality. Consequently, from this perspective alternative JIS appear to be less important than mainstream JIS. For this reason, some researchers send their "best" contributions to WoS or Scopus and their "second best" papers to alternative JIS-indexed journals because they see less value in the latter. These papers add to the number of documents covered by alternative JIS, contributing to their growth, but are perceived as having less worth than those published in mainstream JIS.

Knowledge-bridging

By knowledge-bridging we mean that publishing in alternative JIS provides a link between articles covered by mainstream JIS and "local" communities with limited or no access to it. Mainstream JIS articles are published in journals based in the UK, the USA or the Netherlands, written in English, and generally require payment for access. From the examples in the interviews (such as the use of business and management papers in the classroom, or the linguistic differences between Scielo and WoS in passion fruit publications), we made a list of specific mechanisms through which knowledge-bridging is achieved:

c) **Knowledge adaptation:** The adaptation of knowledge happens when certain concepts or methods are transformed to fit a different context from the original. The study of business history in Latin America, for instance, is conducted through the adaptation of the concepts of business history in high-income countries to low and middle-income countries. Later in its development that adaptation resulted in a differentiated discipline called Latin American business history.

d) **Knowledge diffusion**: Knowledge diffusion occurs, for instance, when a concept that is not novel in mainstream JIS-indexed journals is introduced into a region and shared within the regional community. This can incentivise research on that subject in the regions, as was the case of the introduction of 16S ribosomal RNA sequencing to Colombia.

e) **Teaching:** it is mainly the use of research in alternative JIS for teaching or learning-related activities, as in the introduction of PhD students to academic publications. For instance, papers in alternative JIS-indexed journals help to support and expand the content of their lectures. Lecturers use their own research in articles published in alternative JIS-indexed journals to teach students and use it in their modules.

f) **Business model conversion:** this happens when a researcher publishes open access papers that incorporate bibliographic references from paid journals. For instance, in certain documents such as review papers researchers synthesise literature in mainstream JIS-indexed journals and make it available for readers that cannot afford access to mainstream JIS-indexed paid journals.

g) **Multilingual referencing:** this is when researchers publish in non-English languages and incorporate references from journals in English and other languages. By referencing these multilingual sources researchers build on knowledge that can pose linguistic barriers to readers in their language. This is concretely seen in the incorporation of English-language references into research published in Spanish or Portuguese available through RedALyC and Scielo.

In certain occasions, a conjunction of the mechanisms above can stimulate new areas of study. For instance, the bulk of the production on Latin American business history is covered by alternative JIS, as was indicated by an interviewee and further confirmed through database queries. This sub-discipline emerged from personal interactions with American and British researchers on business history, and currently has grown into into a new area of study. It is mainly published in Spanish and the majority of papers circulate in alternative JIS-indexed journals. From this perspective, alternative JIS serve as a bridge to bring closer knowledge produced by perceived distant communities, with the potential to start novel avenues of research.

Knowledge-gap filling

Knowledge-gap filling is the coverage of knowledge that is neglected or not found in WoS. Examples from business and management, agricultural sciences, and chemistry showed that alternative JIS provide a space for the publication of distinctive original research.

h) **Allow the publication of research that is not well covered or not found in WoS-indexed journals**. Examples include: research that is context-dependent such as Latin American business history or the conceptualisation and application of alternative indicators to understand innovation in countries with low patenting and R&D activity; distinctive subjects such as the production of passion fruit, and research on diseases affecting the production of oil palm; and certain disciplinary areas that have been displaced by others, such as the case of botany that has become less popular than pharmacognosy in high impact factor WoS-indexed journals.

The knowledge gaps that alternative JIS are fulfilling appear to be particularly important in all subjects in which local knowledge ("local" at various scales: from very localised to national to regional) is relevant for policy, management or industrial applications. For example, Arbeláez-Cortés (2013) documented that publications in alternative JIS play a major role in mapping Colombia's biodiversity – an important topic given the country's ecological wealth. In summary, alternative JIS offer a place for the publication of scientific knowledge beyond the boundaries of WoS and Scopus-indexed journals.

DISCUSSION AND CONCLUSIONS

In this article, we have examined the reasons why Colombian researchers publish in journals not indexed by mainstream JIS, i.e. WoS and Scopus. We have found that "lack of scientific quality" of their manuscripts is insufficient explanation to publication patterns. Instead, we have found that knowledge-gap filling, knowledge-bridging, and training towards mainstream JIS-indexed journals were the drivers for publishing in journals in alternative JIS. We believe that the reasons reported in Colombia are likely to apply to other countries in Ibero-America, as well as other developing countries. The extent to which they may also be relevant in other regions for certain topics and disciplines would need to be ascertained.

It follows from these findings that scientometric indicators based on mainstream JIS underrepresent some types of contributions of research from developing countries – as they do not value some training, knowledge-bridging and gap-filling functions. Therefore, research evaluations in Ibero-America should also consider publications in alternative JIS if they wish to value these other types of contributions, which may be particularly relevant in developing countries such as Colombia or other 'peripheral' contexts –i.e. in non-English contexts, for knowledge exchange with non-academic experts or for unconventional topics (Vessuri et al., 2014).

Besides the policy implications, the findings also relate to the theoretical discussion on the universalistic versus particularistic conceptualisations of science. Improvement of the scientific quality partly explains the *training* function of alternative JIS. Since lack of scientific quality is the perception of insufficient research competence as judged by global peers, this is a property that belongs to a universalistic conceptualisation of science. Hence, this publication behaviour can be partly explained by a Mertonian, universalistic conceptualisation of science as an institution.

However, researchers also publish in journals in alternative JIS in order to fulfil *knowledge-bridging* and *gap-filling* functions. Such publishing behaviour does not respond to a universalistic model of science, but to the recognition that mainstream JIS are particularistic institutions, with a lower coverage of journals from developing countries or non-English languages. Our findings thus support the view that scientific institutions such as bibliometric databases are located in specific contexts thus produce a representation of science from a specific, i.e. particularistic perspective. The value of using alternative JIS is to provide different particularistic perspectives of the scientific production, which may be valuable when the evaluation emphasis lies on situated and societal contributions of science (e.g. *gap-filling* and *knowledge-bridging*).

REFERENCES

Arbeláez-Cortés, E. (2013). Knowledge of Colombian biodiversity: published and indexed. *Biodiversity and Conservation*, *22*(12), 2875–2906. http://doi.org/10.1007/s10531-013-0560-y

Braun, V., & Clarke, V. (2006). Using thematic analysis in psychology. *Qualitative Research in Psychology*, *3*(2), 77–101. http://doi.org/10.1191/1478088706qp063oa

Garfield, E., 1996. Significant scientific literature appears in a small core of journals. *The Scientist*, 10(17), 13. Retrieved from http://www.the-scientist.com/?articles.view/articleNo/18038/title/The-Significant-Scientific-Literature-Appears-In-A-Small-Core-Of-Journals/

Lemarchand, G. A. (2012). The long-term dynamics of co-authorship scientific networks: Iberoamerican countries (1973–2010). *Research Policy*, *41*(2), 291–305. http://doi.org/10.1016/j.respol.2011.10.009

Ordóñez-Matamoros, H. G., Cozzens, S. E., & Garcia, M. (2010). International Co-Authorship and Research Team Performance in Colombia. *Review of Policy Research*, *27*(4), 415–431.

Vessuri, H., Guédon J.-C. and Cetto, A.M. (2014) Excellence or quality? Impact of the current competition regime on science and scientific publishing in Latin America and its implications for development. *Current Sociology 62(5)*, 647-665. http://doi.org/10.1177/0011392113512839

Identifying Sources of Scientific Knowledge: classifying non-source items in the WoS

Calero-Medina, C.M.

clara@cwts.leidenuniv.nl
Centre for Science and Technology Studies (CWTS), Leiden University, PO Box 905, Leiden, 2300AK (The Netherlands)

INTRODUCTION

The sources of scientific knowledge can be tracked using the references in scientific publications. For instance, the publications from the scientific journals covered by the Web of Science database (WoS) contain references to publications for which an indexed source record exist in the WoS (source items) or to references for which an indexed source record does not exist in the WoS (non-source items). The classification of the non-source items is the main objective of the work in progress presented here.

Some other scholars have classified and identified non-source items with different purposes (e.g. Butler & Visser (2006); Liseé, Larivière & Archambault (2008); Nerderhof, van Leeuwen & van Raan (2010); Hicks & Wang (2013); Boyack & Klavans (2014)). But these studies are focused in specific source types, fields or set of papers. The work presented here is much broader in terms of the number of publications, source types and fields.

DATA COLLECTION AND METHOD

The first step was to identify the non-source items collected by the WoS[1]. In order to do so we just identified all the non-source items that appear on the references made by the articles and review articles published between 1980 and 2014 on the WoS. The set contains 297,904,154 distinct rows (the unique code number of the citing publication is included)

The information that appears per item in each paper may contains (it is not always the case) information at the level of *Author*, *Volume*, *Issue*, *Page Number*, and a string with *Other information* that may be filled with the title and/or the source. Table 1 shows some examples of non-source items in the WoS.

[1] The Web of Science (WoS) versions of the Science Citation Index and associated citation indices: the Science Citation Index (SCI), the Social Science Citation Index (SSCI), and the Arts & Humanities Citation Index (A&HCI); here the CWTS database containing these records as well as enhanced citation data is briefly indicated as CI. It is important to indicate that the conference proceedings database within the WoS database is not included in this study.

Table 1. Examples of Non-source Items in the WoS.

Author	Year	Volume	Page Number	Issue	Other Information
Fitzgerald, FS	1925				Great Gatsby
Descartes, R					Correspondance
Papavero, N	1978				Catalogue Diptera Am
*Austr Bur Stat	2001				Cens Pop Hous Cdata
Sorenson, DS	1995		1024		P 10 Ieee Puls Pow C
Ducange					Glossarium Mediae In
	1970	1	132		Encyclopedia Polymer
Caves, RE	1989	2	1225		Hdb Ind Org
Gui, MC					In Press J Therm Spr
*Off Nat Stat	2010				Ann Surv Hours Earn
Finney, GH	1975				Thesis Queens U King
Goodwillie, TG		3			Unpub Calculus
Perlez, J	1991		49		NY Times 0922
	1887		2		Figaro 0203
Puccini, G					Madama Butterfly

The non-source item has been identified and classified depending on the source type. As other studies have argued (Nerderhof, van Leeuwen & van Raan (2010) and Boyak & Klavans (2014)) the information at the level of Year, Volume, Issue and Page Number combined with the other fields can be used to help to estimate the type of source. For instance, the cases where all these four fields are filled may be considered Non-Scource Journal or Non-Source Paper. The source type Non-Source Journal/Non-Source Paper has been the first in being identify, since it constitute the largest amount in the dataset. After that searching for keywords and key terms (in different languages) in *Other Information* has helped to identify Conference papers, Handbooks/Manuals, Thesis, Encyclpedia, Survey, In press, Preprint…In the case of Newspapers a more specific strategy has been followed looking for the main newspapers in different countries and then include the names (and possible variables) as search strategies. In the case of Reports (governmental and non-governmental reports) the information at the author level has been helpful since the non-governamental and governamental organizations are identified with an '*'.

In the case of the Books a semi-automatic process has been followed. We have combined the information at the level of *Other Information* and *Author* and search in fields where books are one of the main scientific output and therefore one of the main sources of knowledge as previous studies have shown (e.g. (Hicks (2009), Nerderhof, van Leeuwen & van Raan (2010)). A drawback of this approach is that Scientific and non-Scientific books have not been differentiated.

Table 2 shows how the Non-source Items from Table 1 have been classified following the method briefly explained above.

Table 1. Examples of Non-source Items in the WoS classified by Source Type

Author	Year	Volume	Page number	Iss	Other Information	Source Type
Fitzgerald, FS	1925				Great Gatsby	Book
Descartes, R					Correspondance	Book
Papavero, N	1978				Catalogue Diptera Am	Catalogue
*Austr Bur Stat	2001				Cens Pop Hous Cdata	Census
Sorenson, DS	1995		1024		P 10 Ieee Puls Pow C	Conference Paper
Ducange					Glossarium Mediae In	Glossary
	1970	1	132		Encyclopedia Polymer	Encyclopedia
Caves, RE	1989	2	1225		Hdb Ind Org	Handbook
Gui, MC					In Press J Therm Spr	IN PRESS
*Off Nat Stat	2010				Ann Surv Hours Earn	Survey
Finney, GH	1975				Thesis Queens U King	Thesis
Goodwillie, TG		3			Unpub Calculus	Unpublished
Perlez, J	1991		49		Ny Times 0922	Newspaper
	1887		2		Figaro 0203	Newspaper
Puccini, G					Madama Butterfly	Music

PRELIMINARY RESULTS AND FOLLOW UP RESEARCH

Around two thirds of the initial data non-source dataset have been classified. Overall we have identified 44 source types (see Table 3 below). The frequency of appearances varies greatly but having a refine source type will be of great help for future analysis. The most frequent ones are Non-Source Journals and Non-Source Papers. Under this category are many papers published in journals that are actually covered by the WoS but they are from volumes previous to 1980[2] and papers published after 1980 in Journals only partially covered by the WoS. Conference Papers and Books are also very frequent. Newspapers and Magazines constitute also quite frequent types.

[2] The study is based on the WoS database with publications from 1980 onwards.

Table 3. Source Types Identified

• Archive	• Magazine
• Blog	• Meeting
• Book	• Monograph
• Case Report	• Music Piece
• Catalogue	• Newsletter
• Cd Rom	• Newspaper
• Census	• Non-Source Journal/Non-Source Paper
• Cited indirectly	• Performance
• Communication(*)	• Picture/Illustration/Painting
• Conference Proceedings	• Preprints
• Dictionary/Vocabulary/Thesaurus/Lexicon/Glossary	• Report
• Documentary	• Seminar/Workshop
• Ejournal	• Statistics
• Encyclopedia	• Survey
• Festival Related	• Thesis
• Film/Movie	• Twitter
• Gazette	• Unpublished
• Handbook/Manual	• Video
• In Press	• Website
• Int Tables Cryst	• Wikipedia
• Interview	• Working Paper
• Journal_Periodical	• Youtube

The two main next steps previous to create the final version of the Non-Source Database are:

- For each of the Source Types, select a sample to check the validity of the Source type assignments in order to provide reliable estimates on the validity of our assessments. This is especially important for the Conference Papers, Books and Reports.
- Select a sample from the data that could not be classified yet to learn if some of them could be classified.

Additionally, there will be an attempt to reclassify books in scientific and non-scientific using some mapping and clustering techniques with the help of the VOSviewer software (van Eck & Waltman, 2010).

REFERENCES

Butler, L. & Visser, M.S. (2006). Extending citation analysis to non-source items. *Sicentometrics*, 66, 327-343.

Hicks, D. (1999). The difficulty of achieving full coverage of international social science literature and the bibliometric consequesnces. *Scientometrics*, 44(2), 193-215.

Hicks, D. & Wang, J. (2013). The *New York Times* as a Resource for Mode 2. *Science, Technology & Human Values*, 38(6), 851-877.

Lisée, C., Larivière, V. & Archambault, E. (2008). Conference Proceeding as a Source of Scientific Information: A Bibliometric Analysis. *Journal of the American Society for Information Science and Technology*, 59(11), 1776-1784.

Nederhof, A.J., van Leeuwen, T.N. & van Raan, A.F.J. (2010). Highly cited non-journal publications in political science, economics, and psychology: A first exploration. *Scientometrics*, 83, 363-374.

Van Eck, N.J. & Waltman, L. (2010). Software survey: VOSviewer, a computer program for bibliometric mapping. Scientometrics, 84(2), 523-538.

Data quality and consistency in Scopus and Web of Science in their indexing of Czech Journals

Pavel Mika,* Jakub Szarzec** and Gunnar Sivertsen***

* mika@knav.cz
Library of the Czech Academy of Sciences,
Národní 1009/3, 115 22 Praha 1, Czech Republic

** jakub.szarzec@techlib.cz
National Library of Technology,
Technická 6/2710, 160 80 Praha 6 – Dejvice, Czech Republic

*** gunnar.sivertsen@nifu.no
Nordic Institute for Studies in Innovation, Research and Education (NIFU)
P.O. Box 2815 Tøyen, N-0608 Oslo, Norway

ABSTRACT
This study addresses the discussion of "quality versus coverage" that often arises if a choice is needed between Scopus and Web of Science (WoS). We present a new methodology to detect problems in the quality of indexing procedures. Our preliminary findings indicate the same degree and types of errors in Scopus and WoS. The more serious errors seem to occur in the indexing of cited references, not in the recording of traditional metadata.

Keywords
Scopus; Web of Science; data quality; journal coverage; citation indexing; references; Levenshtein distance.

Submission type: Research in progress paper.
Relevant track: Data infrastructure for research metrics.

INTRODUCTION

This study addresses the discussion of "quality versus coverage" that often arises if a choice is needed between Scopus and Web of Science (WoS). With regard to coverage of source documents and citations, there are large differences in favour of Scopus, although there is not full overlap with WoS content (Gavel & Iselid 2008). The consequences of different coverage depend on the purpose of a particular usage. The two data sources need to supplement each other from an information retrieval perspective (Bar-Ilan 2010). They can, however, replace each other as the basis for indicators of scientific production and citations at the country level (Archambault et al. 2009), but less so at the level of institutions (Vieira & Gomes 2009) or in fields of research that tend to be marginally covered in both sources (Bartol et al. 2014; Haddow & Genoni 2010; Sivertsen 2014).

The quality and consistency of citation indexing procedures are important for all purposes, however. Franceschini et al. (2015) recently published indications of serious types of errors in Scopus that WoS is not free from either. Our study aims at resolving the same question of data quality. We present a new methodology to reach this aim. Our preliminary findings indicate fewer errors and less difference in this respect between Scopus and WoS than we expected from the earlier study. More serious errors seem to occur in the indexing of cited references, not in the recording of traditional metadata. Our further research – also to be presented at the conference – will clarify the extent of this problem.

METHODS

We chose to study journals published by organizations or publishers in the Czech Republic. The reason for this choice is that we wanted to compare Scopus and WoS mainly where they differ: in coverage of the "periphery" of the international core journals. We chose the Czech Republic because the printed versions of the indexed journals are easily available to us. There are 49 Czech journals in the 2014 edition of the Journal Citation Report (WoS) and 159 Czech journals in the 2014 Scopus Journal Title List. Among these, 46 journals are indexed in both databases. They cover Agriculture, Chemistry, Business Economics, Engineering, Plant Sciences, Food Science Technology, Veterinary Sciences, Entomology, Physiology and Microbiology. Most of them (84 per cent) are published in the English language; some are bilingual; the remaining few publish in the Czech language only.

We downloaded the data manually in early December 2015 using the web interface of each database. The queries were limited by ISSN for five years, 2010-2014. We retrieved 13,281 records from Scopus and 13,947 records from WoS in the same 46 journals. The completeness of both downloads was checked against the online versions of the databases after download.

Matching supposedly identical records was crucial in the preparation of data for further analysis. We used an iterative process in several phases where we combined manual and automatic methods based on the Levenshtein distance metric. We were able to match a total of 12,494 records. The matched records thereby constituted 94 percent of the records retrieved from Scopus and 90 percent of the records retrieved from WoS.

The quality and consistency of the data in the two databases was studied by making two types of systematic comparisons. First, the matched records were compared to each other to study possible differences in indexing between the two databases. Second, all records, including those that could not be matched, were compared to the electronic archives of the indexed journals. In addition, two of the journals were analysed using their printed versions. In both types of comparisons, the official indexing policies of the two databases (Scopus Elsevier 2016; Thomson Reuters 2016), which are not identical, provided important guidelines with regard to expected outcomes.

RESULTS

The results of the comparison of the 12,494 matched records are shown in Table 1.

Table 1. Comparison of selected fields

WoS abrevation/name of field		Scopus name of field	Number of identical (provided) values	Base for percentage	% of identical (provided)	Comparison method
AU	Authors	Authors	12,405	12,494	99.3	Number
TI	Document Title	Title	8,394	12,383	67.8	Levensthein
DT	Document Type	Document Type	11,424	12,318	92.7	Identical YES/NO
TC	Times Cited	Cited by	3,713	12,494	29.7	Number
PY	Year Published	Year	12,452	12,494	99.7	Identical YES/NO
VL	Volume	Volume	12,325	12,494	98.4	Identical YES/NO
IS	Issue	Issue	11,766	12,494	94.1	Identical YES/NO
BP	Beginning Page	Page start	12,302	12,494	98.4	Identical YES/NO
EP	Ending Page	Page end	11,944	12,494	95.6	Identical YES/NO
DI	Digital Object Identifier (DOI)	DOI	2,235	2,296	97.3	Identical YES/NO
LA	Language	Language of Original Doc.	11,186	12,494	89.5	Identical YES/NO
DE	Author Keywords	Author Keywords	12,015	12,494	96.1	Number
AB	Abstract	Abstract	11,901	12,494	95.3	Provided YES/NO
NR	Cited Reference Count	Reference count	3,376	4,445	76.0	Number

Generally, we find a high degree of consistency in indexing between the two databases, measured as the percentage identical data in each field, with one important exception, the number of references. All smaller or larger differences between the two databases can be technically explained without altering the general impression that the metadata are of relatively high quality in both databases. Here are several explanations before we turn a discussion of the exception:

- A higher rate of identical titles (68%) could not be expected, because 20 percent of the Scopus titles are multilingual. Other differences were caused mainly by the transcription of technical terms using the Greek alphabet into Latin, for Scopus titles.
- The number of times cited is expected to be different because the two databases cover different numbers of source journals.
- The differences in document type classification are mainly explained because the two resources use different classification schemes. The differences are small. The most common differences are shown in Table 2.

Table 2. Document type differences

WoS doc. type	Scopus doc. type	Number of docs.	% of explored dataset
Article	Review	208	1.7
Review	Article	205	1.7
Proceedings Paper	Article	171	1.4
Editorial Material	Article	139	1.1
Article	Proceedings Paper	107	0.9
Editorial Material	Review	17	0.1

Document type information is important in bibliometric analysis in order to normalize citation indicators. Our results indicate that this type of information is relatively reliable. However, even more important is the indexing of the reference lists in each document. An exception to the finding that metadata are of high quality is the indication we get as we see that 24 percent of the matched records have different reference counts in Scopus and WoS. This is a clear indication that the reference lists in the source documents are not appropriately or fully indexed.

We found 222 WoS records with more references than in Scopus and 847 Scopus records with more references than in WoS. The number of missing references for each comparison is shown in Table 3. The most common difference (12%) was caused by one missing reference in WoS records.

Table 3. Differences in number of references

Reference difference	Number of records	% of records	Number of missing references
WoS>SC	222	5	-1,913
SC>WoS	847	19	2,005
SUM	1,069	24	92

This observation of differences was the starting point for further research when we tried to compare all references from observed records. Unfortunately we still weren´t able to match all the references to find out any pattern in missing (or excess) references.

In the second part of the study, we compared matched as well as unmatched records (Scopus versus WoS) to the electronic archives of the 46 indexed journals. A total of 17,759 records could be used for the study of how and to what extent the journals are indexed. A quantitative overview is given for each of the journals in Table 4 (Appendix). Here, we compare the number of records in the original source journal to the number of records indexed in WoS and Scopus and the number of records that could be matched between them. No numbers are the same for any of the journals and there are wide differences for some journals. The right column in Table 4 (A-C) refers to the following explanations for the differences:

A. There are only small differences for nine journals. The differences can mainly be explained because of differently defined document types used for indexing hybrid journals with a large array of document types.
B. There are larger differences between Scopus and WoS for 25 journals; however, the number in one of the databases resembles the number of records in the original source. The differences between the two indexing databases can be explained by differing indexing policies, with the exceptions below.
C. There are large differences between the original sources and the two indexing databases for nine journals. In these cases, we found that the electronic archive of the journal does not cover the journal completely or the archive includes supplemental items not published in the regular journal.

An example of C is *Chemicke listy* (0009-2788), where the archive includes supplementary material such as conference abstracts of plenary lectures, oral sessions and posters.

Differences of type B were examined by inspecting the printed versions of two journals. In *Folia Biologica* (ISSN 0015-5500), we discovered that the larger number of records in Scopus was caused by an error in which 71 records from a Polish journal with the same name but different ISSN (0015-5497) were included. We also found two instances of duplicate records in Scopus. All in all, we found 14 cases of the duplicate Scopus records in the whole dataset, which is less than expected from earlier studies of the same error (Valderrama-Zurián et al. 2015).

Inspecting *Československá psychologie* (0009-062X) in the same way, we found that neither Scopus nor WoS covered this journal completely. In spite of the indexing policy, 12 items were not indexed in WoS – mostly news, errata, and discussions. Of 214 items not indexed by Scopus, 51 were classified as research articles in WoS. If this classification is correct, they should have been indexed in Scopus according to its policy. The other missing items in Scopus can be explained by the policy of not indexing such items.

DISCUSSION AND FOCUS FOR FURTHER RESEARCH

We have established a methodology for two types of comparisons that aim to test the quality and consistency of the data and indexing in Scopus and WoS, by:

a. Matching and measuring the degree of similarity in supposedly identical records in both databases.
b. Comparing data from both databases to the sources that were indexed.

With both methods, most of the differences we observed could be explained according to differing methods and policies for indexing in Scopus and WoS or the specific publishing policies of journals.

There are two major exceptions, however, that will be the focus of our further studies:

a. Differences in the number of cited references in a record may be an indication that reference lists in the source documents are not appropriately or fully indexed.
b. Differences between the number of records in the archive of the source journal and the databases can be an indication that the contents are not appropriately or fully indexed.

REFERENCES

Archambault, É, Campbell, D., Gingras, Y., & Larivière, V. (2009). Comparing bibliometric statistics obtained from the Web of Science and Scopus. Journal of the American Society for Information Science and Technology, 60(7), 1320-1326.

Bar-Ilan, J. (2010). Citations to the "Introduction to informetrics" indexed by WOS, Scopus and Google Scholar. Scientometrics, 82(3), 495-506.

Bartol, T., Budimir, G., Dekleva-Smrekar, D., Pusnik, M., & Juznic, P. (2013). Assessment of research fields in Scopus and Web of Science in the view of national research evaluation in Slovenia. Scientometrics, 98(2), 1491-1504.

Franceschini, F., Maisano, D., & Mastrogiacomo, L. (2016). The museum of errors/horrors in Scopus. Journal of Informetrics, 10(1), 174-182.

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

Haddow, G., & Genoni, P. (2010). Citation analysis and peer ranking of Australian social science journals. Scientometrics, 85(2), 471-487.

Scopus Elsevier. (2016). Content Policy and Selection. (https://www.elsevier.com/solutions/scopus/content/content-policy-and-selection) [21 March 2016].

Sivertsen, G. (2014). Scholarly publication patterns in the social sciences and humanities and their coverage in Scopus and Web of Science. In E. Noyons (Ed.), Proceedings of the science and technology indicators conference 2014 Leiden (pp. 598-604). Leiden: CWTS.

Thomson Reuters. (2016). The Thomson Reuters Journal Selection Process. (http://wokinfo.com/essays/journal-selection-process/) [21 March 2016].

Valderrama-Zurián, J., 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.

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

Table 4. Number of records in the original source journal compared to the number of records indexed in WoS and Scopus, and the number of records that could be matched between them. The right column (A-D) refers to explanations for the differences given in in the text.

Journal title abbrev.	Source N	WoS N	SC N	Matched N	Differences
Acta Ent Mus Nat Pra	293	291	241	239	B
Acta Geodyn Geomater	216	218	217	217	A
Acta Chir Orthop Tr	393	346	360	340	C
Acta Vet Brno	388	398	388	388	B
Agr Econ-Czech	304	302	292	292	B
Appl Math-Czech	176	175	169	169	B
Biomed Pap	728	346	344	317	C
B Geosci	234	233	245	220	B
Cent Eur J Publ Heal	233	224	286	219	B
Ceram-Silikaty	278	280	278	278	A
Cesk Slov Neurol N	668	594	574	536	C
Cesk Psychol	223	378	176	171	B
Czech J Anim Sci	331	320	317	317	C
Czech J Food Sci	408	406	386	386	B
Czech J Genet Plant	208	207	192	191	B
E M Ekon Manag	237	286	234	232	B
Epidemiol Mikrobi Im	193	171	165	132	C
Eur J Entomol	445	409	403	398	C
Financ Uver	136	137	136	130	A
Folia Biol-Prague	199	198	272	198	B
Folia Geobot	144	136	137	135	C
Folia Microbiol	439	439	457	437	B
Folia Parasit	237	232	211	211	B
Folia Zool	198	193	194	191	A
Fottea	106	108	103	103	A
Hortic Sci	128	128	126	126	A
Chem Listy	4,160	1,290	1,254	1,033	C
J Appl Biomed	133	129	106	103	B
J Geosci-Czech	127	126	111	110	B
Kybernetika	344	343	337	335	B
Listy Cukrov Repar	603	416	456	390	C
Morav Geogr Rep	109	86	108	84	B
Neural Netw World	218	212	205	198	A
Photosynthetica	379	373	401	364	B
Physiol Res	631	634	619	610	B
Plant Protect Sci	134	97	132	88	B
Plant Soil Environ	445	437	437	437	A
Polit Ekon	266	279	218	214	B
Prague Econ Pap	136	138	128	127	B
Preslia	127	127	127	120	A
Radioengineering	728	735	725	724	A
Slovo Slovesnost	177	160	73	66	B
Sociol Cas	760	471	204	197	C
Soil Water Res	111	110	111	110	A
Stud Geophys Geod	224	224	232	223	B
Vet Med-Czech	404	405	394	387	B

Missing citations due to exact reference matching: Analysis of a random sample from WoS. Are publications from peripheral countries disadvantaged?[1]

Paul Donner

donner@dzhw.eu
Deutsches Zentrum für Hochschul- und Wissenschaftsforschung, Berlin
Schützenstraße 6a, 10117 Berlin (Germany)

INTRODUCTION

Citation counts of scientific research contributions are one fundamental data in scientometrics. Accuracy and completeness of citation links are therefore crucial data quality issues (Moed, 2005, Ch. 13). However, despite the known flaws of reference matching algorithms, usually no attempts are made to incorporate uncertainty about citation counts into indicators. This study is a step towards that goal. Particular attention is paid to the question whether publications from countries not using basic Latin script are differently affected by missed citations.

The proprietary reference matching procedure of Web of Science (WoS) is based on (near) exact agreement of cited reference data (normalized during processing) to the target papers bibliographical data. Consequently, the procedure has near-optimal precision but incomplete recall - it is known to miss some slightly inaccurate reference links (Olensky, 2015). However, there has been no attempt so far to estimate the rate of missed citations by a principled method for a random sample. For this study a simple random sample of WoS source papers was drawn and it was attempted to find all reference strings of WoS indexed documents that refer to them, in particular inexact matches. The objective is to give a statistical estimate of the proportion of missed citations and to describe the relationship of the number of found citations to the number of missed citations, i.e. the conditional error distribution. The empirical error distribution is statistically analyzed and modelled.

DATA AND METHODS

The analyzed data originate from licensed raw data in tagged format of the WoS journal and proceedings citation indexes. A simple random sample of target items was drawn from all journal articles, letters and reviews, as defined in the data. The WoS citation links were obtained, as given by the WoS matchkey, the T9/R9 fields. All citations until 2015 are counted.

Reference data of all publications from 1980 to spring 2015 are indexed for search, the sampled target source items are from the same period. The reference strings, consisting of author name field, split into last name and initials at the comma, the source title, publication year and first page fields were indexed with Oracle Text. The volume field was not considered because for the target journal items volume and publication year are nearly redundant information but publication year is more accurate and more complete in WoS

[1] This work was supported by BMBF project 01PQ13001. The author would like to thank Anastasiia Tcypina for her help in collecting the data.

reference data than volume (Moed, 2005, table 13.1). A procedure was programmed to search the index for references likely referring to the sampled target items. Because the search field data of the target items has to correspond to the way the reference data is prepared by WoS, the target author name and source title were pre-processed accordingly. This entailed reduction of first name to initials, the removal of non-letter characters in the name, and, for the source title, using the WoS abbreviation. Where more than one abbreviated title version for the same journal existed in the data and when an additional group first author was available, all possible combinations of those fields' values were used as search input. The procedure performs a fuzzy search on the index and returns a list of unique candidate reference strings that are sufficiently similar to the target input. The search is deliberately lenient so that all possible matches are returned in order to prevent false negatives as much as possible, which is a requirement for this study.

The candidates were reviewed clerically on whether they constitute a match to the target or not by a student assistant. Ambiguous candidates were afterwards assessed by the author. Care was taken to avoid false positive matches by querying the database for any exact matches of the candidate reference strings other than the target item. The found positive matches are used as additional citation links and the derived extended citation count for each target item is calculated by retrieving all references using those candidate strings.

PRELIMINARY RESULTS

A total of 372 cases were assessed. The distribution of missed citations per item is presented in table 1.

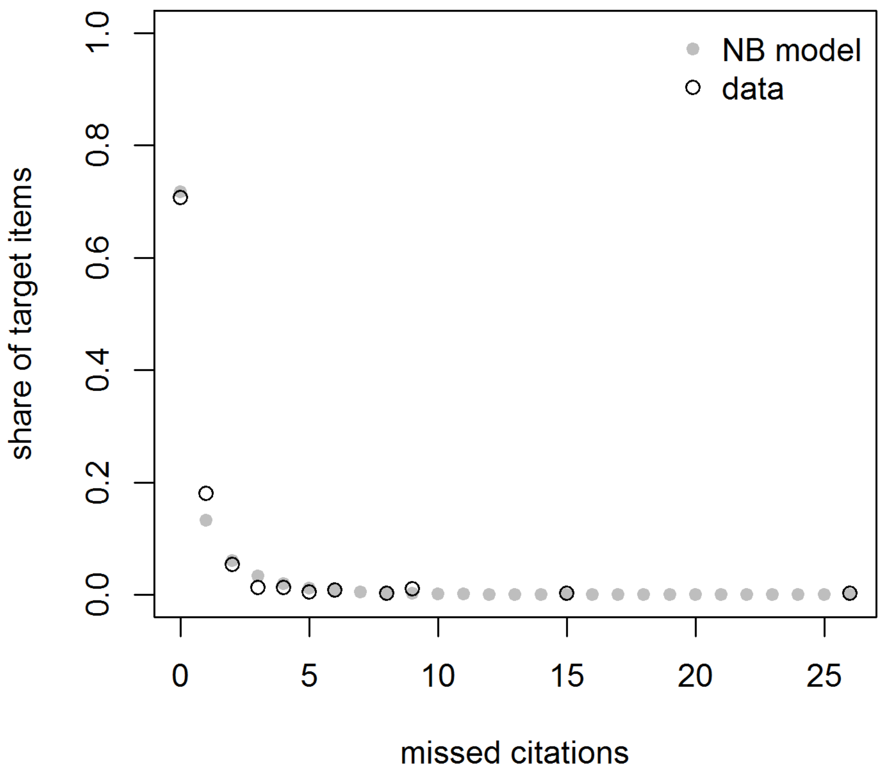
Table 1: Distribution of number of missed citations

missed citations	0	1	2	3	4	5	6	8	9	15	26
occurrence	263	67	20	5	5	2	3	1	4	1	1

29.4 % of the target items have one or more detected missing citation. The average of citations per paper (CPP) according to WoS is 16.4; the average of missing CPP is 0.7. In total, according to WoS' exact citation links, the papers were cited 6120 times. 255 additional citations were found. Thus, 4 % of citations were missed. An association between the apparent citation count and the citation count error can be observed, as the number of WoS citations and the number of missed citation per item are correlated with $r = 0.31$.

Citation distributions can be approximately modelled by negative binomial (NB) distributions with reasonable accuracy (Schubert and Glänzel, 1983; Ajiferuke and Famoye, 2015). As the error distribution is also discrete, non-negative and heavily skewed, it was attempted to model it with the NB distribution. Numerical estimation using the R package *MASS* gave the following parameter estimates: $\theta = 0.254$ (SE = 0.039) and $\mu = 0.685$ (SE = 0.083) and the model fit is depicted in Figure 1.

Figure 1: Empirical probability distribution of the data and first NB model



The goal is not just to describe the distribution in general by a model, but to predict the error, that is, the number of missing citations, for a given number of WoS citations. The above model does not take the association between WoS citation count and missed citations into account, which was shown by the correlation. The model is next extended by regressing the parameter μ on the WoS recorded citation counts w, while holding the scale parameter θ of the binomial regression fixed to the previously estimated value.
The estimated equation is: $\mu = 0.5 + 1.012 \times w$.[2]

Having estimated the parameters, we can now simulate random deviates from this distribution or obtain values of the density or cumulative probability at any desired point. For example, according to the model, the probability of having zero missing citations for a publication with WoS citation count of 0 is 75.9 %, for one missing citation is 12.8 %. For 100 WoS citations, the probability of having 0 missing citations is 21.8 %; the probability for one missing citation is 5.5 %.

By Monte Carlo simulation from the model one may obtain a predicted *distribution* of the sum of missed citations, in this case for 372 publications. In a Bayesian statistical framework this is the posterior probability distribution of the parameter of interest. To do this, we make 372 draws from the model, that is, a random NB variable with the estimated model

[2] SE of the intercept: 0.182. SE of the coefficient of *w*: 0.004; both significant at the 0.01 level.

parameters. To take model uncertainty into account, inputs (the θ and μ parameters) are not fixed but are also randomly drawn from truncated normal distributions (as only values > 0 are possible) with the mean being the point estimate and the standard deviation being the estimated standard error. This means that parameter estimates are replaced by prior distributions with hyperparameters from the preceding estimations.

The sum of the values of an iteration is the estimate of the sum of missed citations. The procedure was replicated 10.000 times to get an approximation of the probability distribution of the number of missed citations, given the model and estimates. The distribution obtained is characterized in table 2:

Table 2: Summary of the simulated distribution of the sum of errors (10.000 replications)

min.	median	mean	max.	credible 95 % interval
131	257	260	492	(195, 341)

Recall that the observed total of missing citations is 255. The model point estimate of missed citations is 260 with a Bayesian credible interval derived from the quantiles of the posterior distribution of [195, 341].

PERIPHERAL EFFECTS

Eastern European and East Asian researchers often encounter ambiguities when they have to transliterate their names in to basic Latin script or when their names are simplified to basic Latin script for database indexing. Because reference matching relies on author names, it is hypothesized that publications from those peripheral regions are subjected to comparatively higher risks of missed citations. To test the hypothesis, all first author country information were coded into three mutually exclusive regions and one category for unknown country. This nominal variable with three levels was added to the second model. Three papers' first authors had both one address of a peripheral region and one of a non-peripheral one. In these cases, they were coded as the peripheral region. The distribution of publications over regions is displayed in table 3. Furthermore, publication year was also added to the model as a predictor to see if any temporal change in reference accuracy can be detected.

Table 3: Regions defined as "peripheral" and countries

Region	Publications	Countries
East Asia	48	China, Japan, South Korea, Taiwan
Eastern Europe	23	Ukraine, Russia, USSR, Serbia, Bulgaria, Czech Republic, Czechoslovakia, Serbia and Montenegro, Hungary, Poland, Latvia
Others	301	all others

The regression equation is

$$\mu = \beta_1 + \beta_2 w + \beta_2 r + \beta_3 p;$$

with r being the variable region, p the publication year and the parameter θ held constant to 0.254 as before.

In the expanded regression, the coefficients for the regions were found to be not significant at the 0.05 level with 'Others' as the reference level. The effect of publication year was not significant either[3]. The coefficient of WoS citation count (β_2) is slightly smaller but remains significant at the 0.01 level. Thus, the hypotheses that the country of the first author or the publication year affects the citation error rate are rejected for this sample. The clear limitation of this study is that the group sizes are so small that differences are difficult to detect, so an extended sample, possibly stratified by region, might reveal contradictory evidence.

References

Ajiferuke, I., & Famoye, F. (2015). Modelling count response variables in informetric studies: Comparison among count, linear, and lognormal regression models. *Journal of Informetrics*, *9*(3), 499-513.

Moed, H. F. (2005). Citation Analysis in Research Evaluation. Springer, Dordrecht, The Netherlands.

Olensky, M. (2015). Data accuracy in bibliometric data sources and its impact on citation matching. Dissertation, Humboldt-Universität zu Berlin.

Schubert, A., & Glänzel, W. (1983). Statistical reliability of comparisons based on the citation impact of scientific publications. *Scientometrics*, *5*(1), 59-73.

[3] The publication year was significant at the 0.1 level.

STI Conference 2016 · València Peripheries, frontiers and beyond
València, 14 · 16 September 2016
21st International Conference on Science and Technology Indicators | València (Spain) | September 14-16, 2016

Funding Acknowledgements in the Web of Science: inconsistencies in data collection and standardization of funding organizations[1]

Jeroen van Honk*, Clara Calero-Medina* and Rodrigo Costas*

*j.s.van.honk@cwts.leidenuniv.nl; clara@cwts.leidenuniv.nl; rcostas@cwts.leidenuniv.nl
Leiden University (Centre for Science and Technology Studies - CWTS), P.O. Box 905, Leiden, 2300 AX (The Netherlands)

ABSTRACT

Funding acknowledgements (FAs), as provided by the Web of Science, are a recent inclusion in the bibliometric toolset. They are starting to be used to study the presence, relationships and effects of funding and funders in the production of new scientific research. The incorporation of this new source of information comes with the need to understand how these data are collected and indexed in the database. This paper discusses important inconsistencies related to the method by which the data for FA and funders are selected, extracted and indexed by WoS, thereby highlighting the need to construct a thesaurus for the data. Problematic areas are found to be the quality of the input data and the conceptualization of what does and does not constitute a funding acknowledgement. Based on this critical analysis of the data and the identification of the main issues, we provide several recommendations for researchers, funders, WoS and other users of the data.

INTRODUCTION

Funding acknowledgements (FAs) have recently started to be included in the bibliometric toolset as a source of information to study the presence, relationships and effects of funding and funders in the production of new scientific research. This has been possible particularly since the Thomson Reuters Web of Science (WoS) database started to collect FA information from August 2008 onwards. The inclusion of this relevant piece of information has opened up new possibilities in the field of acknowledgements research (Costas & van Leeuwen, 2012; Desrochers et al., 2015; Díaz-faes & Bordons, 2014) and particularly in the area of FA studies (Sirtes, 2013). The availability of FAs in scientific publications allows the study of the presence of funding across disciplines, different funders or co-funding patterns in science (Wang & Shapira, 2011). It is important to highlight that in addition to the FA information, Thomson Reuters also collects the full funding text from scientific publications as well as the grant number, if provided in the publication, thus opening the possibility of more refined analysis of specific funding programs.

However, the incorporation of this new source of information also comes with the need to understand how these data are collected and indexed in the database. The importance of knowing the boundaries of the data collected is critical for the adequate use of this new source of information with analytical purposes (Paul-Hus et al, 2016). Previous studies have already pointed out some of the limitations of the FA data collected by WoS. For example, Rigby (2011) pointed out the presence of misspellings of funding bodies or errors in grant numbers, a problem that has also been addressed by Sirtes (2013), which sought to correct the severe

[1] This work has been partially supported by funding from the DST-NRF Centre of Excellence in Scientometrics and STI Policy (South Africa).

problems of data standardization in WoS funding information. Additionally, important limitations on the coverage of these data have recently been reported (Paul-Hus et al, 2016). In Paul-Hus et al (2016) an internal guideline policy document from Thomson Reuters was discussed (Thomson Reuters Bibliographic Policy Funding Acknowledgements, 2015). The existence of these guidelines supports the idea that the FA data collection has an important decision component, where indexers are expected to evaluate and select funding texts and funding bodies for their indexation (or not), thus opening also possibilities of introducing inconsistencies in the selection and indexation of the data.

OBJECTIVES

This paper highlights the challenges that arise in the construction of a thesaurus of funding organizations based on the funding acknowledgement data from the WoS. We will discuss important inconsistencies related to the method by which the data for FA and funders are selected, extracted and indexed by WoS. Thus, the main objective of this paper is to provide a critical analysis on central methodological aspects related with the collection and standardization of FA data in WoS.

METHODOLOGY AND DATA COLLECTION

The WoS source data are collected on three levels, namely (1) the full funding text, (2) the extracted funding body, and (3) the extracted grant number. The funding body and grant number are linked where they occur together.

A cursory look at these funding bodies indexed by WoS makes it abundantly clear that, without any standardization, the quality of data is highly variable. It can be argued that there are two main central issues related to FA data quality: 1) great variation in the acknowledging practices held up by researchers, thus creating a diversity of funding organization names[2], grant numbers, mentions of support, etc.; and 2) the relatively undisclosed and occasionally unclear means by which WoS identifies and collects the FA data. These two obstacles together make the data challenging and specific solutions (e.g. thesauri, contextual analysis of the funding texts) need to be considered in order to develop meaningful analyses from the FA data. The two main issues we encountered in the standardization of FA WoS data can be grouped under: 1) inconsistencies in the selection of FA data to be indexed; and 2) inconsistencies of the FA data indexed in WoS.

Inconsistencies in the selection of FA data to be indexed

It has been reported that WoS only collects FAs when the acknowledgement[3] section of the publication contains funding-related information (cf. Paul-Hus et al, 2016). In our database, barring a negligible number of exceptions[4], there are no funding texts which are not connected to either at least one funding body or grant number. However, the criteria regarding which types of funding information are selected and indexed for which acknowledgements remain unclear. An entity can be indexed as funding body after simply being thanked for

[2] These are caused, among many other things, by spelling mistakes, varying translations, and formatting variances.

[3] Acknowledgement sections generally contain more than just funding acknowledgements, with acknowledging practices tending to extend to more generic expressions of support (Costas & van Leeuwen, 2012; Díaz-faes & Bordons, 2014).

[4] A total of only 104 cases, most of them from 2009, when the data were just starting to be collected by WoS.

"support", without further specification as to the form this support took.[5] Whether such unspecified cases actually concern financial, technical, intellectual, or material support is not always evident, even when taking into account the rest of the funding text.

Those funding bodies that are extracted are not standardized. The method seems to be restricted to a simple tagging of phrases occurring in the text. While such a method seems intuitive enough, it proves to be problematic in its distinction between implicit and explicit funding acknowledgements. Consider the following pair of examples:

"This study was supported by internal funding from UNC Health Care".

"This study was supported by internal funding sources".

Of these two, the first text is indexed, with "UNC Health Care" listed as funding body. The second is not indexed. In order to be consistent, these cases ought to be treated equivalently. Another example can be the following:

"*The authors wish to thank the National Institute of Malaria Research (Indian Council of Medical Research), Delhi, India for encouragement and moral support.*".

It is unclear why this acknowledgement is indexed (and the "National Institute of Malaria Research (Indian Council of Medical Research), Delhi, India" extracted as funder). Obviously this is not a FA and the support mentioned is "moral", not financial.

It is important to remark that the study of this type of inconsistency is very complex as it requires an analysis of which acknowledgements have been selected for indexing and which have not. We are currently working on a more extensive analysis to explore further these limitations (van Honk, Calero-Medina, & Costas, 2016).

Inconsistencies in the indexation of funding bodies

Like the decision of what constitutes a funding text, the extraction of funding bodies from these texts is not always self-evident. One problem arising in the WoS FA data is when two funding entities have been incorrectly lumped together and presented as one (i.e. "NSF/DOE" and "National Science Foundation/Department of Energy", 346 occurrences in the WoS data). The reverse also occurs: two funding agencies identified separately while they are actually parts of the same whole. This happens for instance with the "Program for Changjiang Scholars and Innovative Research Team in University (PCSIRT)" string, where the part before and after "and" are occasionally (yet not consistently) indexed separately. The lack of consistency is notable: in otherwise similar circumstances one acknowledgement is split while the other is kept whole.[6] A similar inconsistency is found for the "U.S. EPA's Science to Achieve Results" scholarship, which is sometimes split in "U.S. EPA's Science" and "Achieve Results" (this happened in at least 85 cases). These examples suggest that a manual examination on a paper-by-paper basis is bound to introduce inconsistencies in the data

[5] i.e. "BP would like to thank ICRANet for support on this project." ICRANet is in fact an Italian research institute, and this unqualified "support" could have taken many forms, yet the institute is indexed as a funding body by WoS.

[6] Within acknowledgement texts containing the phrase "Changjiang Scholars and Innovative Research Team in University", there are many more indexations for the second part of the conjunction (474) than for the first (86).

extraction and parsing, especially when what constitutes a "funding acknowledgement" seems to have been loosely conceptualized by WoS.

The clue to another important inconsistency is in the occurrence as funding bodies of organizations that are not strictly funding organizations (e.g. universities acknowledged and indexed by WoS as funders). It turns out that when universities occur as FA they usually occur as the recipient of funding rather than as funder. Take for instance the following funding text:

"This material is based upon work supported by the National Science Foundation under Grants Nos. CHE-0721505 and CHE-0809053 at the University of Arizona (SGK) and under Grant No. CHE-719157 at the Massachusetts Institute of Technology. Cossairt and Cummins would also like to thank Thermphos International for support."

From this text, WoS has extracted three funders: "National Science Foundation at the University of Arizona", "Massachusetts Institute of Technology" and "Thermphos International". Here, two universities are indexed as funding bodies while in reality only one funding body is mentioned here: the National Science Foundation.[7]

Issues related with the indexation of grant numbers

The availability of grant number information becomes very helpful to identify and resolve some of the issues detailed above. Grant numbers provide a far more structured and patterned source of data, although when they appear as serial number only they are naturally ambiguous. Moreover, though more structured, grant numbers nevertheless pose data problems of their own, particularly in the variations in which they appear (e.g. EY014801 also appears as: "NEI P30 Core Grant EY014801", "NEI P30 EY014801", "P30 EY014801", "P30 EY-014801", "P30EY014801", "P30-EY014801").

It is the frequent presence within the grant numbers of alphabetical characters, hyphens and other forms of punctuation which makes it possible to identify highly specific grant number patterns belonging to specific funding organizations, thus allowing the possibility to assign the acknowledgements in which these numbers are mentioned to these organizations. The US National Institutes of Health provide a good example of this. Their grant numbers clearly signify the individual institutes. For example, the strings "EY" (National Eye Institute) or "CA" (National Cancer Institute) provide a way to identify these institutes as funders even when they are not explicitly mentioned in the FA.

CWTS thesaurus

Considering all the inconsistencies mentioned above, CWTS has commenced with the creation of a thesaurus of funding organizations and sources extracted by WoS. The construction of this thesaurus follows a similar approach to the cleaning of affiliation data carried out for the Leiden Ranking (Waltman et al., 2012). Thus, a large extent of the FA data provided by WoS has been mapped to thesaurus entries, as such creating a new, cleaner and more workable data set, including funding organizations (Wellcome Trust, Deutsche Forschungsgemeinschaft); funding schemes (Cancer Center Support Grants, Horizon 2020), and organizations mentioned in the FA that are not primarily funding-oriented (e.g.

[7] The extraction of Thermphos International (a former venture in the chemical industry sector) as funding body also provides another example of indexation on the basis of an unspecified, general "support".

universities, research institutions, etc.). In the process, hierarchical connections between funding bodies have also been established, so that for instance the National Cancer Institute (US) has been identified as a child institute of the National Institutes of Health (US). The leading rule in cleaning up the data has been to retain as much of the information inherent in the data as possible, but up to a certain threshold.[8] If a funding body does not reach this threshold, its acknowledgements are mapped to its parent agency (if possible). The goal here is to strike a balance between richness and usability of data.

Overall, in the CWTS thesaurus, more than 450 funding organizations, 230 funding schemes, and 6400 organizations have been identified and cleaned from the FA data indexed by WoS.

The thesaurus merges the funding body and grant number data, based on a set of rules (generally preferring grant number data in case of conflict). See the examples below:

Table 1. Thesaurus.

WoS funding body	WoS grant number	CWTS Thesaurus
Univ Michigan	013448-001	Univ Michigan - Ann Arbor
Univ Michigan	--	Univ Michigan - Ann Arbor
--	R01 CA122443	U.S. National Cancer Institute
Univ Michigan	R25 CA112383	U.S. National Cancer Institute

As this table shows, by taking data from both funding sources our Thesaurus becomes both more complete and more accurate.

DISCUSSION

The introduction of new FA data by WoS has opened up important possibilities of expanding the bibliometric toolset to the study of acknowledgements in general and funding information in particular. This study shows however that there are important aspects that need to be considered for the proper use and understanding of the FA data collected by WoS.

In the first place, the strong variation in the data collected and indexed by WoS requires the development of specific automated algorithms (e.g. Sirtes, 2013) and thesauri in order to be able to perform reliable studies on standardized lists of funding organizations.

Secondly, this study has shown the presence of important inconsistencies in the selection and indexation of FA data. The baseline of such a problem is that it is not straightforward what a “funding acknowledgement” actually is. The inconsistency between different types of support declarations and their consequent indexation shows that a more robust discussion on what is considered a “FA”, its typologies and its theoretical and conceptual operationalization, is still lacking. Therefore, the inconsistencies found in this paper are relatively unsurprising, since this lack of conceptualization has a direct effect on the operationalization of the term. This also raises the question if it should be the role of a data provider such as WoS to decide which FA to collect or not. In this regard, it would be much more helpful if WoS was to focus on collecting all acknowledgements from scientific publications (without deciding whether these constitute FAs or not), index all entities that appear in them, and simply leave to the

[8] A funding body needs a minimum of 500 acknowledgements, currently.

bibliometricians and expert analysts the delineation and conceptualization of what they consider a FA.[9]

Thirdly, this study also highlights important inconsistencies in the indexation of the entities that are funding the research and those that are funded. The availability of grant numbers has proven to be a good instrument to clarify and correct potential mistakes. However, this opens questions on what is currently being indexed as "funders" in WoS. Again, this also calls for a more thorough and conceptual model of what is a FA and what are the actors involved in them (e.g. which is funder and which is funded).

Based on this study, several recommendations can be proposed for several stakeholders: researchers, funders, WoS and users of FA data:

1. Funders should provide the funded authors with clear funding statements containing a uniquely identifiable, explicitly mentioned, and standardized form of the funder name (including where possible grant numbers and other distinctive codes). Funders should also inform authors on how to clearly disclose the type of support (e.g. economic, travel, access to resources, etc.) they have received, allowing the possibility of better narrowing down on the types of support related to the funder.

2. WoS would rather strive towards being comprehensive in collecting all acknowledgements from scientific publications and extracting all entities mentioned therein, without making a priori decisions on what constitutes a FA. This would clearly contribute to a more consistent database of acknowledgements and acknowledged entities, while also expanding the bibliometric scope of their data by opening the possibility to study all types of acknowledgements (Cronin, McKenzie, & Stiffler, 1992).

3. Scientometric researchers and practitioners need to observe caution when working with FA WoS data, as not all "funding bodies" and acknowledgements that are collected by WoS necessarily constitute funders, and some of the FA metadata may also have some inaccuracies, omissions and deficiencies, making their use problematic. In addition, it is important to count with thesauri or standardized methodologies in order to be able to properly work with the FA data provided by TR (in a similar fashion as it was necessary for WoS affiliation data, cf. Fernández, et al. (1993)).

[9] This would also open the possibility to studying what distinguishes "internal" from "external" funding; what represents a conflict of interest disclosure, and which specific types of support (e.g. access to materials or equipment, travel support) constitute "funding".

REFERENCES

Costas, R., & van Leeuwen, T. (2012). Approaching the "Reward Triangle": General Analysis of the Presence of Funding Acknowledgments and " Peer Interactive Communication" in Scientific Publications. *Journal of the American Society for Information Science and Technology*, 63(8), 1647–1661. doi:10.1002/asi.22692

Cronin, B., McKenzie, G., & Stiffler, M. (1992). Patterns of acknowledgment. *The Journal of Documentation*, 48(2), 107–122.

Desrochers, N., Paul-Hus, A., Bowman, T., Costas, R., Haustein, S., Larivière, V., … Tsou, A. (2015). Authorship, patents, citations, acknowledgments, tweets, reader counts and the multifaceted reward system of science. In *Proceedings of the American Society for Information Science and Technology*. Retrieved March 29, 2016 from https://www.asist.org/files/meetings/am15/proceedings/openpage15.html

Díaz-faes, A. A., & Bordons, M. (2014). Acknowledgments in Scientific Publications □: Presence in Spanish Science and Text Patterns Across Disciplines. *Journal of the Association for Information Science and Technology*, 65(9), 1834–1849. doi:10.1002/asi.23081

Fernández, M. T. T., Cabrero, A., Zulueta, M. A. A., & Gómez, I. (1993). Constructing a relational database for bibliometric analysis. Research Evaluation, 3(1), 55–62.

Honk, J. van, Calero-Medina, C., & Costas, R. (2016). A case study on the accuracy of Web of Science funding acknowledgement data: PLoS. In preparation.

Paul-Hus, A., Desrochers, N.; & Costas, R. (2016). Characterization and description of funding acknowledgement data in Web of Science: key issues in their consideration as a research tool. Submitted to *Scientometrics*.

Rigby, J. (2011). Systematic grant and funding body acknowledgement data for publications: new dimensions and new controversies for research policy and evaluation. *Research Evaluation,* 20(5), 365–375.

Sirtes, D. (2013). Funding Acknowledgments for the German Research Foundation (DFG). The dirty data of the Web of Science database and how to clean it up. In J. Gorraiz, E. Schiebel, C. Gumpenberger, M. Hörlesberger, & H. Moed (Eds.), *Proceedings of the 14th International Society of Scientometrics and Informetrics Conference* (pp. 784–795). Vienna: AIT GmbH.

Thomson Reuters. (2015). Thomson Reuters Bibliographic Policy Funding Acknowledgements. Internal document.

Waltman, L., Calero-Medina, C., Kosten, J., Noyons, E. C. M., Tijssen, R. J. W., Eck, N. J. Van, … Wouters, P. (2012). The Leiden Ranking 2011 / 2012 □: Data Collection , Indicators , and Interpretation. *Journal of the American Society for Information Science and Technology*, 63(12), 2419–2432. doi:10.1002/asi.22708

Wang, J., & Shapira, P. (2011). Funding acknowledgement analysis: an enhanced tool to investigate research sponsorship impacts: the case of nanotechnology. *Scientometrics*, 87(3), 563-586.

Open Data in Global Environmental Research: Findings from the Community[1]

Birgit Schmidt*, Birgit Gemeinholzer and Andrew Treolar

*bschmidt@sub.uni-goettingen.de
State and University Library, University of Göttingen, Platz der Göttinger Sieben, Göttingen, 37070 (Germany)

INTRODUCTION

This paper presents findings from the Belmont Forum's survey on Open Data which targeted the global environmental research and data infrastructure community (Schmidt, Gemeinholzer & Treloar, 2016). It highlights users' perceptions of the term "open data", expectations of infrastructure functionalities, and barriers and enablers for the sharing of data. A wide range of good practice examples was pointed out by the respondents which demonstrates a substantial uptake of data sharing through e-infrastructures and a further need for enhancement and consolidation. Among all policy responses, funder policies seem to be the most important motivator. This supports the conclusion that stronger mandates will strengthen the case for data sharing.

The Belmont Forum, a group of high-level representatives from major funding agencies across the globe, coordinates funding for collaborative research to address the challenges and opportunities of global environmental change. In particular, the E-Infrastructure and Data Management Collaborative Research Action has brought together domain scientists, computer and information scientists, legal scholars, social scientists, and other experts from more than 14 countries to establish recommendations on how the Belmont Forum can implement a more coordinated, holistic, and sustainable approach to the funding and support of global environmental change research.

METHODS

In the context of the working group on Open Data researchers from various science communities, interested laypersons, government employees, and others were invited to share their views and experiences on data publishing, access and (re)use. The main aim of the survey was to learn more about:

- Key open data activities in various communities dealing with global environmental change to identify leading examples of best practice from a user perspective;
- Areas where users' desire to share could be enhanced by new/other developments;
- Barriers to "open data sharing" from a user perspective (as either a data provider or data user).

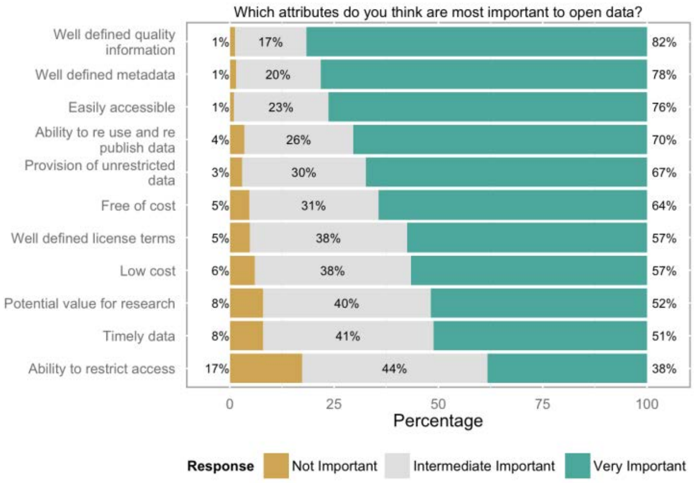
From September to November 2014, the survey collected over 1,300 responses based on the distribution of the survey to about 20 disciplinary and professional mailing lists, and to all the authors of a well-renowned open access publisher, central to the research area. All of the 19 questions of the survey were non-mandatory. For the analysis the statistics software R and in particular the Likert package were used. All data are available via the Zenodo repository (Schmidt et al, 2016).

[1] This work was supported by the Belmont Forum's E-Infrastructure and Data Management Research Action.

KEY FINDINGS

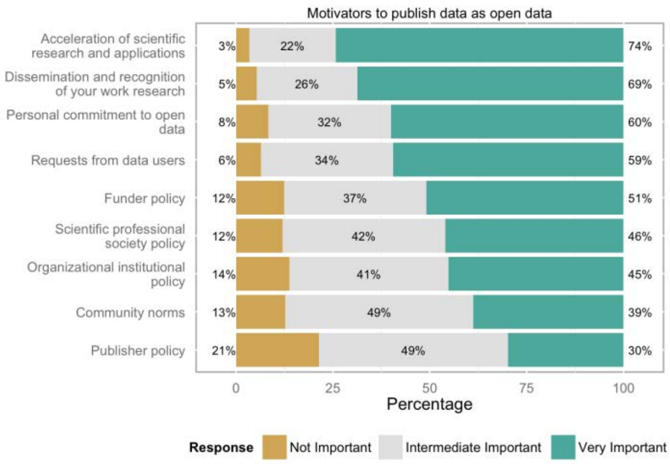
Instead of providing a definition of "open data" the survey assessed the user perception associated with the term (compare Fig. 1). The answers highlight the importance of information which enables the user to assess the quality of data, to select data based on metadata, and to easily access and reuse the data. The ability to restrict access was lowest in the ranking of desirable attributes, which fits the intuitive idea of openness. However, nearly 2/5 off all respondents still considered the option to restrict data as a very important attribute.

Figure 1: Perceived properties of open data.



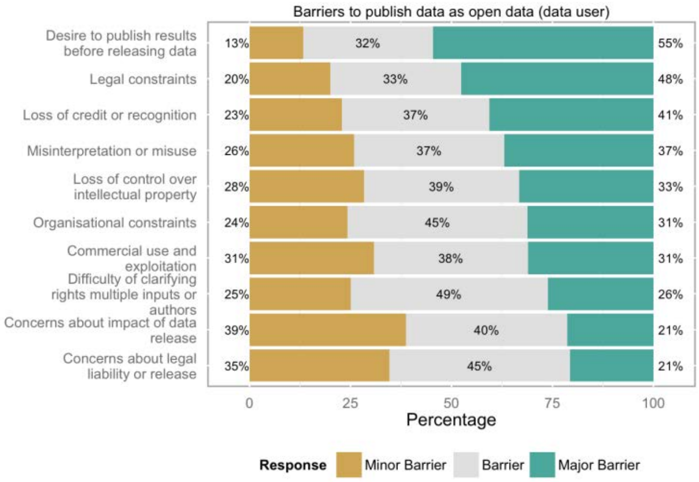
Motivators and barriers to publish data as open data were studied in the survey. The main desires to publish data as open data arose from research-intrinsic motives ranging from general considerations, i.e. the acceleration of scientific research and applications, to personal motivations, i.e. dissemination and recognition of research results, personal commitment to open data and requests from data users (cf. Fig. 2). Among the three types of data professionals which responded to the survey (data user, data provider, data manager) data managers' personal commitment to open data seem to be significantly higher.

Figure 2: Motivators to publish data as open data.



Overall, the most important barrier for publishing data as open data were the desire to publish results before releasing data, legal constraints, loss of credit or recognition and possible misinterpretation or misuse. Concerns about legal liability for data or release of data were least pronounced. In addition, the desire to publish results before releasing data was somewhat more prevalent at early stages of a research career.

Figure 3: Barriers to publish data as open data.



In addition, the survey explored where the community accesses and/or publishes data, and a wide range of good practice examples was pointed out by the respondents (several of these data repositories are currently added to the re3data.org registry) which demonstrate a substantial uptake of data sharing and reuse through data e-infrastructures in the global environmental change community. A need for further enhancement and consolidation can be derived from the respondents' expectations about functionalities of infrastructures and desires expressed about access to specific types of data.

CONCLUSIONS

Based on the findings of the survey the following actions were recommended to the Belmont Forum:

- Funders should make open data archiving mandatory, while taking into account the main motivators revealed by the survey.
- Scientific merit as well as accelerating research and applications are still the main motivators for publishing data; thus ethics of data sharing and reuse should be taken into account when proposing guidelines for open data sharing and re-use.
- Support and training activities should be supported in concerted ways, targeting researchers as well as current and future data and information professionals.
- Interoperability between infrastructures should be further facilitated, taking into account generic requirements (e.g. providing links to publications and funder information) as well as disciplinary norms and standards (e.g. vocabularies, metadata standards).

REFERENCES

Schmidt, B., Gemeinholzer, B. & Treloar, A. (2016). Open Data in Global Environmental Research: The Belmont Forum's Open Data Survey. *PLoS ONE*, 11(1), e0146695. doi:10.1371/journal.pone.0146695.

Schmidt, B., Gemeinholzer, B., Treolar, A., Hodge, J., Santanchè, A. & Oakley, K. (2015). Belmot Forum Open Data Survey. Zenodo, doi:10.5281/zenodo.16384

Examining data access and use in science1

Erjia Yan*, Mengnan Zhao*

*ey86@drexel.edu; mz438@drexel.edu
College of Computing and Informatics, Drexel University, 3141 Chestnut Street, Philadelphia, PA (U.S.A.)

ABSTRACT
In this research-in-progress paper, we provide preliminary evidence of data access and use in scientific literature based on a content analysis of 600 stratified sampled PLOS ONE publications. Results show that data access and use varied greatly from one paper to another in terms of how datasets were collected, referenced, and curated.

INTRODUCTION
Data have progressively become an integral component in modern science—thanks to the increased permeability of disciplinary boundaries, the enhanced human mobility, and the advanced technologies to process, analyze, and curate large scientific data. Scientists can now form interdisciplinary, international collaborative teams surrounded by data to conduct the so called data-centric or data intensive research (Tansley & Tolle, 2009).

In science, there is a growing awareness of data access and sharing. As early as 2004, scholars have advocated for an international framework to promote data accessibility (Arzberger et al., 2004). It is argued that data sharing helps develop a democratic society (Harrison et al, 2012), enhances the transparency of scientific research particular for those sensitive topics such as climate change (Parmesan & Yohe, 2003), allows for reproducing and validating research (Bradley, 2009; Nosek et al., 2015), and unleashes the potential of data to solve complex societal issues such as diabetes (Zimmet et al., 2001). Realizing these benefits, a number of scientific journals and funding agencies have begun mandating making data freely available to the public: for instance, *Nature* requires authors to "make materials, data, code, and associated protocols promptly available to readers without undue qualifications" (Nature Editor, n.d.), and likewise the National Science Foundation of the U.S. expects investigators to "share with other researchers, at no more than incremental cost and within a reasonable time, the primary data, samples, physical collections and other supporting materials created or gathered in the course of work under NSF grants" (National Science Foundation, n.d.). Organizations have also made an effort of indexing data such as Thomson Reuters' *Data Citation Index* (Thomson Reuters, n.d.) or SageCite by University of Bath, U.K. (Lyon, 2010).

Despite these efforts, access to data is still highly inconsistent and even obscure. Data can be formally curated in journal-specific digital repositories or institutional archives that are typically assigned with DOIs or URLs, or informally stored in personal computers and

[1] This work was made possible in part by the Institute of Museum and Library Services (Grant Award Number: RE-07-15-0060-15), for the project titled "Building an entity-based research framework to enhance digital services on knowledge discovery and delivery". In addition, this work was also supported by the National Consortium for Data Science (NCDS) for the Data Fellows program.

servers. As a result, data are referenced in unsystematically ways in scientific literature: they can be formally cited, or simply be mentioned in paragraphs, footnotes, endnotes, and acknowledgements. A 2014 study on oceanography data access found that data are more likely to be mentioned in the text than been formally cited (Belter, 2014). Thus, merely using citation indices is insufficient to capture the different ways of data access and use. Instead, full-text publications provide the crucial context for this purpose. This research-in-progress paper reports a preliminary set of results on several key aspects of data access and use in science.

METHODS

DATA SOURCE

The data set used in the study contained open access, full-text papers from PLOS ONE. The access to the data set is provided by the PubMed Central Open Access Subset (http://www.ncbi.nlm.nih.gov/pmc/tools/openftlist/) and is publically available. We sampled 50 papers randomly from 12 defined discipline for papers published during 2014-2015, which resulted in 600 papers in total.

CODING PROCEDURES

We employed content analysis as the research instrument because it is an effective method to discover quantitative patterns from textual corpora (Bauer, 2000; Herring, 2010; Krippendorff, 2012). In content analysis, coding is the crucial link between data collection and data interpretation, allowing researchers to use a set of guidelines (i.e., coding schemes) to systematically make sense of data.

We first created a draft coding scheme and then adopted the grounded theory approach and applied the draft scheme to a subset of the data with the goal to identify previously unnoticed yet valuable patterns—this process helped us complement the coding scheme and the finalized version is shown in Table 2. Most of the coding items are pre-coordinated, while new codes may emerge during coding, which are referred to as emergent codes. We marked the emergent codes with "*" and kept refining the coding scheme during the whole coding process. The final coding scheme is shown in Figure 1. It should be noted that while most PLOS ONE publications may have used data to certain extent (i.e., quantitative research), in this study, we operationalized data as datasets—data that were stored in certain formats or media (for instance, a paper that used a statistical analysis without explicitly mentioning datasets is therefore not considered as a paper that used data in research). Because of the unambiguous and self-explanatory nature of the coding scheme as well as an obtained inter-rater reliability of 1 on a small sample of the data, one coder coded all 600 papers.

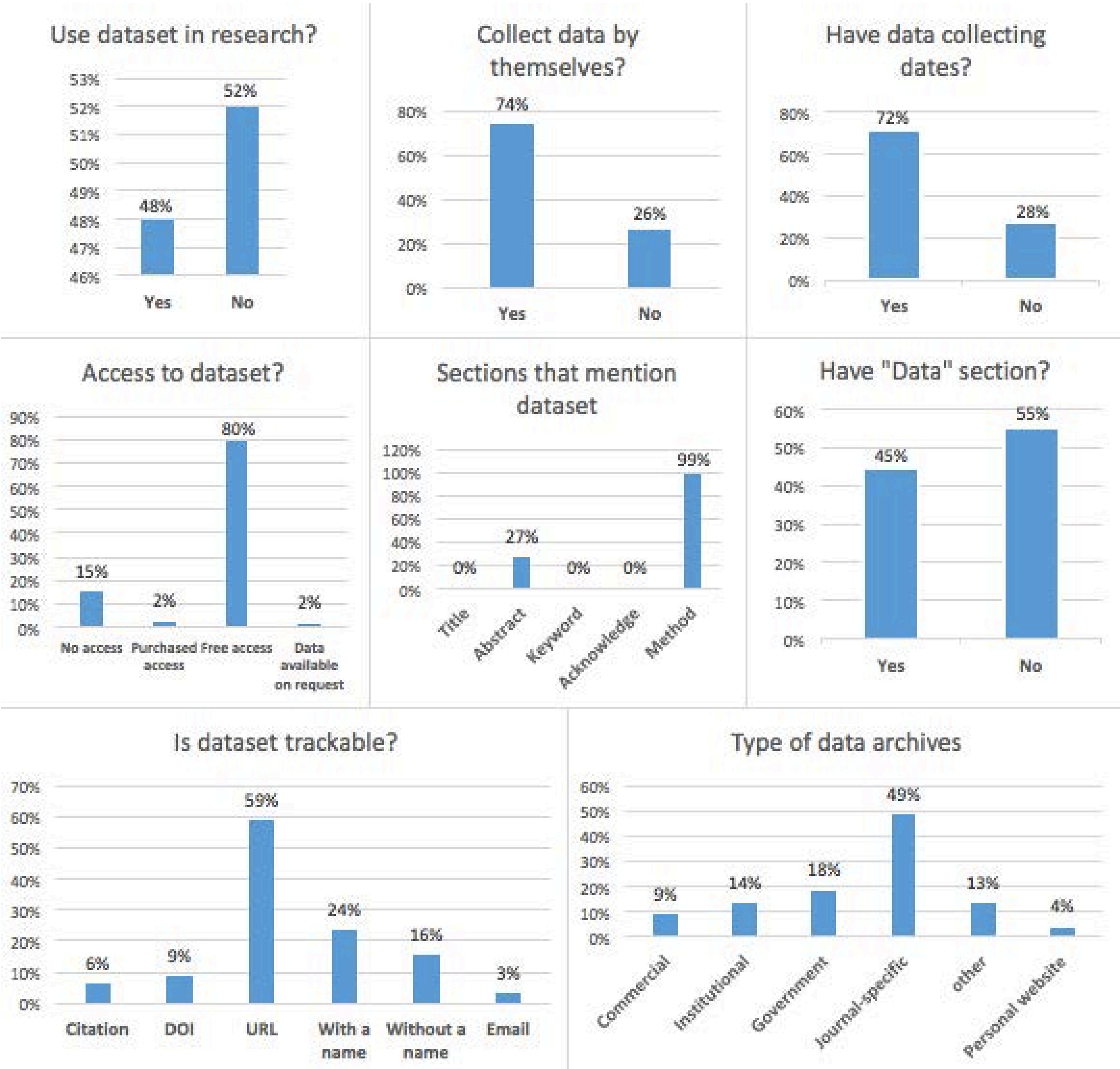
Figure 1. Coding Scheme

1. Use data in research	4. Data access	6. Section of data mention
1.1 Y	4.1 No access	6.1 Title
1.2 N	4.2 Purchased access	6.2 Abstract
2. Data section	4.3 Free access	6.3 Keyword
2.1 Y	*4.4 Data available on request	6.4 Acknowledge
2.2 N	5. Data tracking	6.5 Method
3. Data collection	5.1 Citation	7. Type of archives
3.1 Collecting data on their own	5.2 DOI	7.1 Commercial
3.1.1 Data collection date	5.3 URL	7.2 Institutional
3.1.1.1 Y	5.4 With a name	7.3 Governmental
3.1.1.2 N	5.5 Without a name	7.4 Journal-specific
3.2 Using public data set	*5.6 Email	7.5 Other
		*7.6 Personal website

RESULTS

After finishing the coding work, we went through the coded articles and counted the number of articles in each coding category, with the results shown in Figure 2.

Figure 2. Percentage of articles in each coding category



Among the 600 articles in the data set, 52% (or 312 articles) used datasets in their research. Within the 312 articles that used datasets, only about half of them had data or data related sections; the others just mentioned data sources, data collection methods, or data descriptions in method sections. Within the 312 articles, 74% collected data and created datasets by themselves; among these, 72% provided the date or the time period when data were collected. The numbers demonstrate that a majority of studies are inclined to create datasets and use their own data, rather than reuse previously created and curated data from others' research.

For data attribution methods, citations and data identifiers are considered the most normative by facilitating ways of data tracking; however, only 6% and 9% of the articles respectively attributed data in such formal ways. Meanwhile, 60% of the articles provided URL to locate datasets. Most of the provided URLs worked at the present time, but the concern is that once an URL expires, we will lose track of the datasets. Furthermore, 24% articles just provided the names of the datasets, some of which are quite unique to be located while others may refer to multiple data entities. Some articles also attached an email address with the dataset so that readers can send email requests for data.

In regards to means of data storage, nearly half of them saved data in the journal website as attachments to papers, followed by housing data in governmental (18%), institutional (14%) and commercial (9%) repositories. In addition, 4% articles hosted data in researchers' personal websites.

CONCLUSIONS

In this research-in-progress paper, we provided preliminary evidence of data access and use in scientific literature based on a content analysis of 600 stratified sampled PLOS ONE publications. Results showed that data access and use varied greatly from one paper to another in terms of how datasets were collected, referenced, and curated. The next step in this research project will involve the identification of disciplinary characteristic of data access and use as well as the design of inclusive indicators to comprehensively capture the full-spectrum of data impact.

REFERENCES

Arzberger, P., Schroeder, P., Beaulieu, A., & Bowker, G. (2004). An international framework to promote access to data. Science, 303(5665), 1777.

Bauer, M. (2000). Classical content analysis: A review. In M. Bauer & G. Gaskell (eds.), *Qualitative Researching with Text, Image and Sound* (pp. 131-151). Thousand Oaks, CA: Sage.

Belter, C. W. (2014). Measuring the value of research data: a citation analysis of oceanographic data sets. PloS one, 9(3), e92590.

Bradley, J. C. (2007). Open notebook science using blogs and wikis. Retrieved June 1, 2016 from http://precedings.nature.com/documents/39/version/1

Harrison, T. M., Guerrero, S., Burke, G. B., Cook, M., Cresswell, A., Helbig, N., ... & Pardo, T. (2012). Open government and e-government: Democratic challenges from a public value perspective. Information Polity, 17(2), 83-97.

Herring, S. C. (2010). Web content analysis: Expanding the paradigm. In J. Hunsinger, M. Allen, & L. Klastrup (Eds.), *The International Handbook of Internet Research* (pp. 233-249). Berlin: Springer Verlag.

Krippendorff, K. (2012). Content analysis: An introduction to its methodology. Sage.

Lyon, L. (2010). UK Digital Curation Centre: enabling research data management at the coalface. Microsoft Research Faculty Summit 2010, Redmond, Washington, USA, July 2010.

National Science Foundation (n.d.). Dissemination and Sharing of Research Results. Retrieved March 10, 2016 from https://www.nsf.gov/bfa/dias/policy/dmp.jsp

Nature editor (n.d.). Availability of data, material and methods. Retrieved March 10, 2016 from http://www.nature.com/authors/policies/availability.html

Nosek, B. A., Alter, G., Banks, G. C., Borsboom, D., Bowman, S. D., Breckler, S. J., ... & Contestabile, M. (2015). Promoting an open research culture: Author guidelines for journals could help to promote transparency, openness, and reproducibility. Science (New York, NY), 348(6242), 1422.

Parmesan, C., & Yohe, G. (2003). A globally coherent fingerprint of climate change impacts across natural systems. Nature, 421(6918), 37-42.

Tansley, S., & Tolle, K. M. (Eds.). (2009). The fourth paradigm: data-intensive scientific discovery (Vol. 1). Redmond, WA: Microsoft research.

Thomson Reuters (n.d.). Data Citation Index. Retrieved March 10, 2016 from http://wokinfo.com/products_tools/multidisciplinary/dci/

Zimmet, P., Alberti, K. G. M. M., & Shaw, J. (2001). Global and societal implications of the diabetes epidemic. Nature, 414(6865), 782-787.

SMS: a linked open data infrastructure for science and innovation studies*

Peter van den Besselaar*, Ali Khalili**, Al Idrissou*, Antonis Loizou**, Stefan Schlobach**, Frank van Harmelen**

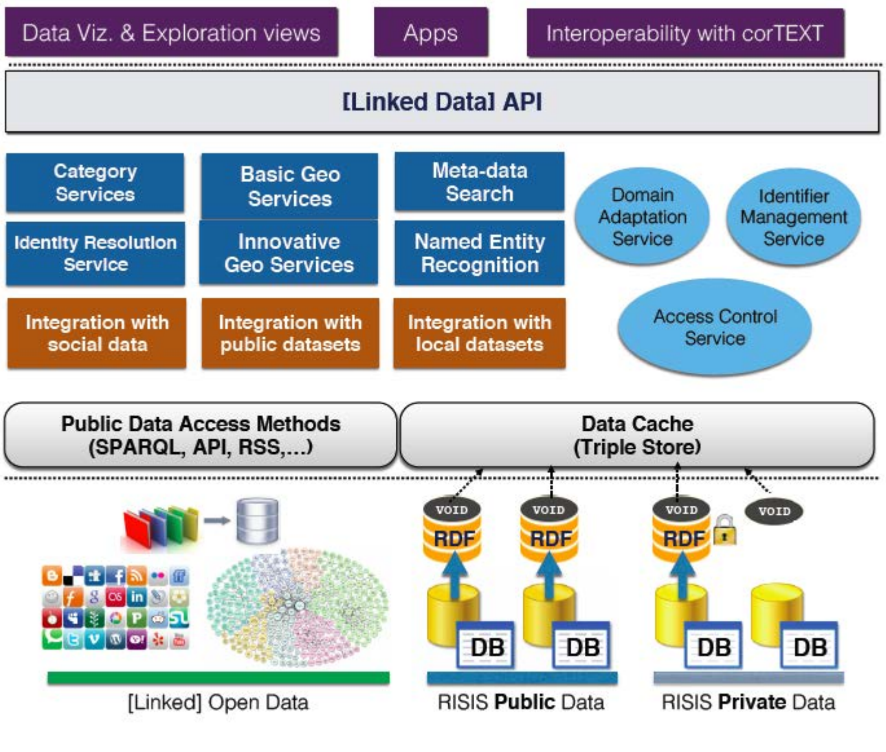
p.a.a.vanden.besselaar@vu.nl
* Network Institute & Department of Organization Sciences
Vrije Universiteit Amsterdam, Amsterdam, The Netherlands

** Network Institute & Department of Computer Science
Vrije Universiteit Amsterdam, Amsterdam, The Netherlands

ABSTRACT
In this paper we describe a data integration infrastructure for Science Technology and Innovation (STI) studies developed within the context of the RISIS project. We outline its architecture and functionalities. In the full paper, we will show the use of the infrastructure in a complex research project. At the conference we will give a demonstration.

INTRODUCTION
In this paper we describe a data integration platform for science technology and innovation data using semantic web technology and focusing on (but not restricted to) linked open data (Beek et al 2016). Figure 1 shows the basic architecture.

Figure 1: SMS architecture



* EC grant 313082, the RISIS project: Research Infrastructure for Science and Innovation Studies.

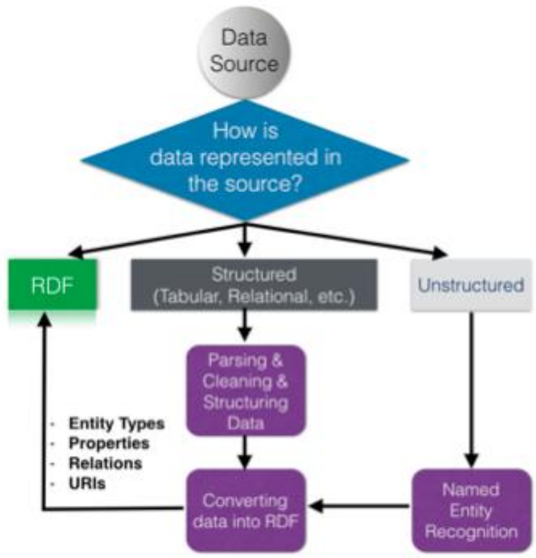
Why is such infrastructure needed? Up to now, STI studies are either *rich* but small scale (qualitative case studies) or large scale and *under-complex* – because they generally use only a single dataset like Patstat, Scopus, WoS, OECD STI indicators, etc., and therefore deploying only a few variables – determined by the data available. However, progress in the STI research field depends in our view on the ability to do large-scale studies with often many variables specified by relevant theories: There is a need for studies which are at the same time big *and* rich. To enable that, combining and integration of STI data and beyond is needed – in order to exploit the many data that are „out there" in an innovative and meaningful way. That is why the core of the infrastructure is the conversion of different datasets in the same open format: from tabular data, text data and web data to RDF (Beek et al 2016).

This emphasis on data integration is also visible in other research fields. That enables to build a data infrastructure partly by reusing existing tools. Within the RISIS project we develop the *SMS platform for data integration and data enrichment* by combing those existing tools with specific tools newly developed for the STI field. The SMS platform partly implemented now; we aim at providing a complete beta version later on this year, as part of the RISIS S&I data infrastructure (www.risis.eu). The following functions can be distinguished in the SMS platform:

Pre-processing

Pre-processing data and converting data into the RDF standard for linked open data (Figure 2). For example, PDF files can be converted into TXT, and through *Named Entity Recognition* relevant entities like people, organizations, countries, etc. are identified. Additional text processing (e.g., term extraction) may identify attributes. A concrete example is recognizing research institutions and universities in a researcher"s CV, using name recognition by linking the CV to databases with background knowledge such as DBpedia. The resulting data are then converted into RDF. Structured data (e.g., Excel files) are parsed and cleaned. And then converted into RDF.

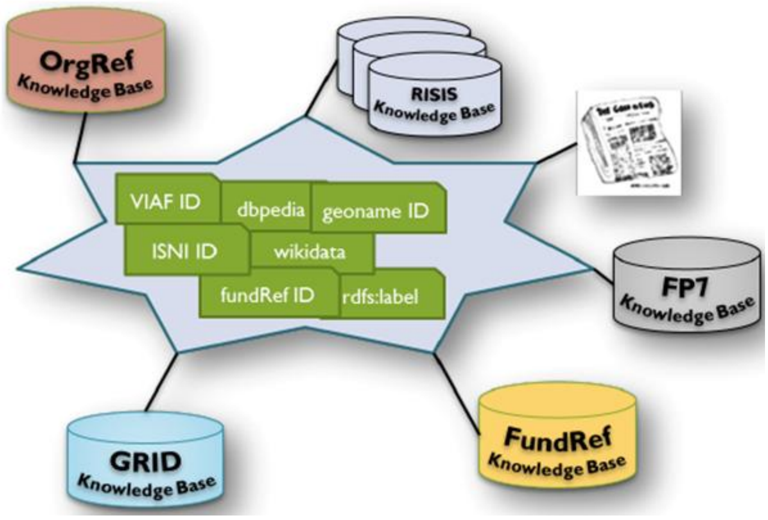
Figure 2: From heterogeneous data to RDF



Linking data

The next step is *linking* the data. If entity *identifiers* are available, the linking is easy. If not, a variety of techniques can be used, from (fuzzy) *string matching* to *deploying attributes* available in the different databases. If names occur in different languages, resources like DBpedia can be used to match. If two entities have different names, but similar characteristics, they may be in fact the same entity. However, whether entities are considered the same, depends on the perspective: sometimes two organizations (e.g. departments) can be the same – because they are parts of the same organization (university). But if one wants to compare departments, this is not the case. We are currently experimenting with a series of datasets on research organizations, in order to compile basic reference sets of research organizations. This is done through interlinking different datasets through knowledge resources on the web (Figure 3).

Figure 3: Linking data through web knowledge resources



There is also the problem of disambiguation of person names, which is (in our field) mainly related to publications and patents, and for which specific tools are being developed (Sandström & Sandström 2009; Gurney et all 2012). One of the questions addressed is how complete disambiguation needs to be given the questions posed in a research project (Van den Besselaar & Sandström, forthcoming). Within the SMS platform we do not tackle this problem field for the moment, and the researcher can use existing tools – together with still quite some manual work.

Geo-services

An interesting possibility is linking through geo-location: if two entities have the same geo-location, they may be related (or identical). Geo-locating has an additional advantage, as it is also an instrument to enrich data: many other (open) data provide variables measures at some level of geographical aggregation: e.g., environmental data, educational data, or socio-economic data. Therefore the platform provides a variety of *geo-services.*

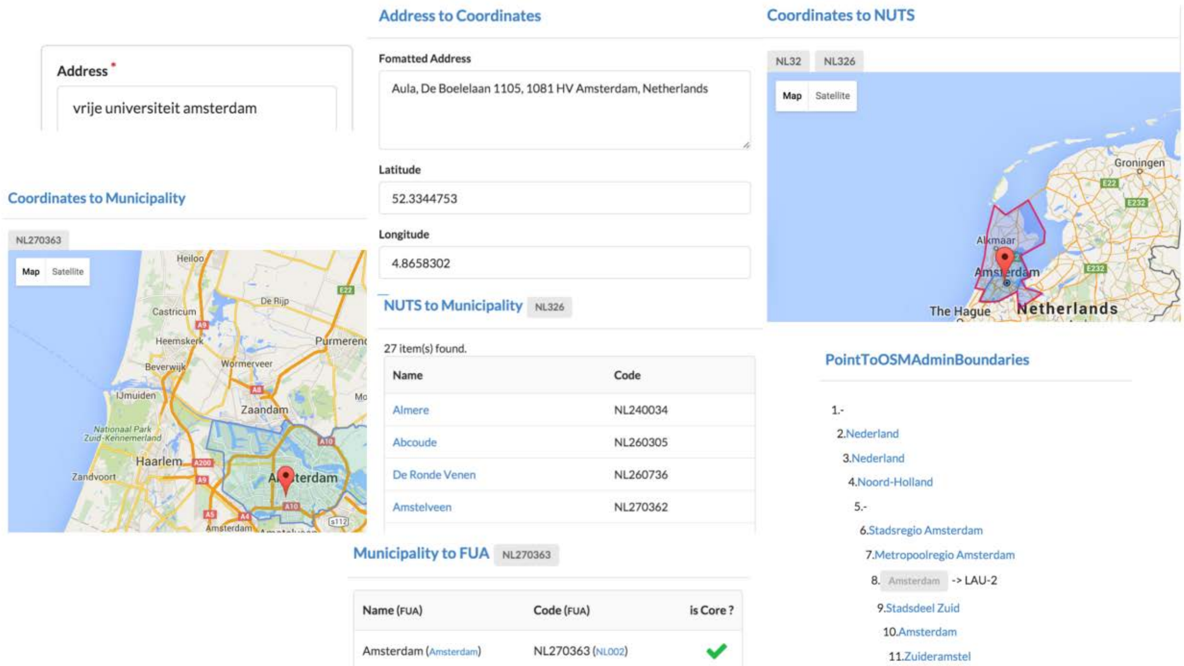
We illustrate this with an example of a service to determine the geographical location if one knows an address (or even only a name). The system is based on a series of open geo-resources, such as GeoNames and OpenStreetMap (figure 4). In the top left part of the screen the address "Vrije Universiteit Amsterdam" is inserted, and the service has as output various maps and, in the bottom right, the geo-characterization of the inserted address at eleven levels. Figure 4 shows the various administrative boundaries where Level 8 represents LAU 2.

By integrating these resources, the service can give for an entity‟s address the geo-location at 11 different levels, which then can be used to link the entity to other (often statistical) data. Of course the platform can be used to do this for larger amounts of addresses, and the output then is not on the screen, but in a tabular form. In the future we aim at adding different distance concepts, such as travel distance (time, frequency, price, etc.).

Figure 4: Geo-locating services

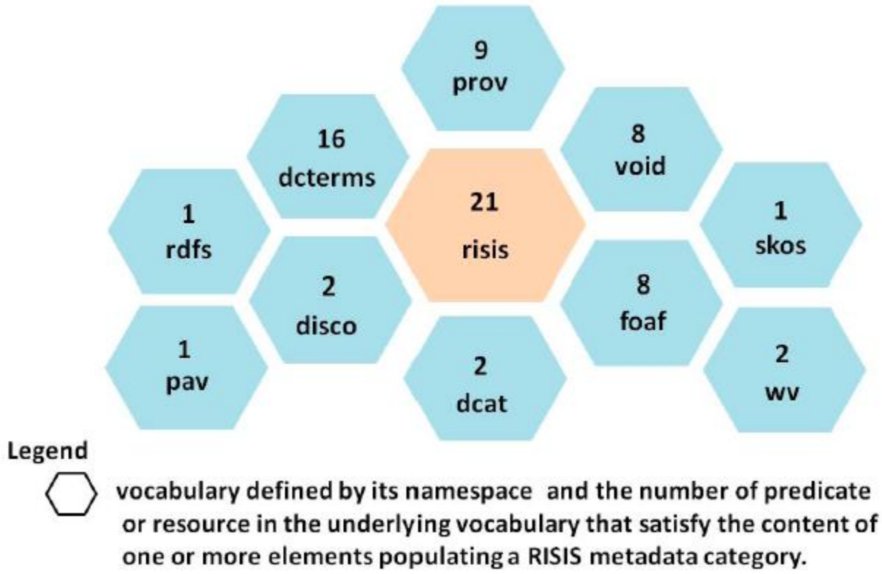
Category services

As datasets may use different category systems for the attributes, linking data requires a mapping of these category systems or „vocabularies‟ (Figure 5). A good example are the different systems that are used for classifying research fields, e.g., in the Web of Science and in OECD R&D statistics. A *category service* would enable the data user to select which one



classification he/she wants to use. And the system would then do the mapping between the different classifications. For this, we deploy existing vocabularies available on the web. One can also think of other classification schemes that can be mapped, e.g., of professions, of jobs, of types of organizations, and so on. As many developments are taking place, the SMS platform may use what is available. E.g., within the RISIS project work is done on classifications of companies, and of research organizations. The RISIS metadata system will be of help here.

Figure 5: Integrating vocabularies



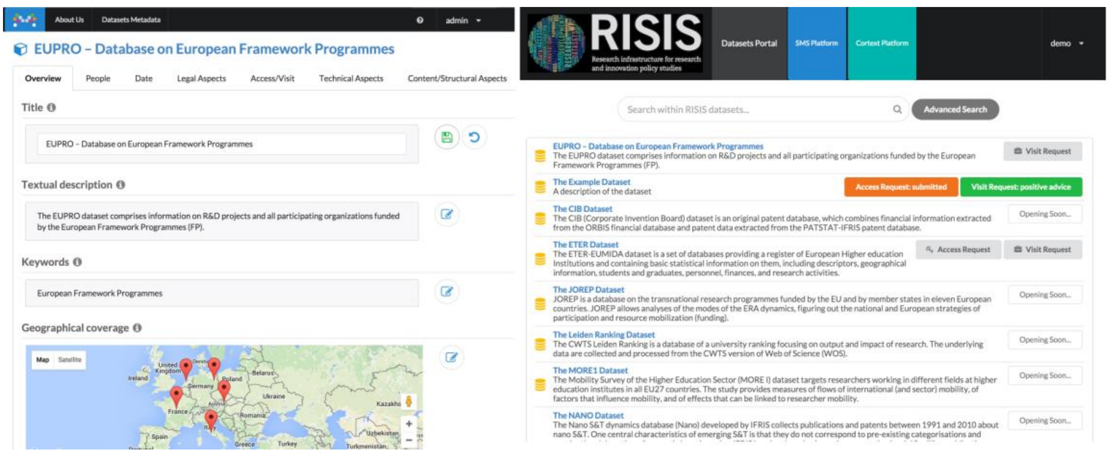
Improving quality and data enrichment

Linking can also be used to *improve quality* of the data and *enrich* them. Linking the two sets may increase the number of variables, but also may reveal discrepancies in variable values, and the user should then be able to decide what the more reliable source is. Quality improvement follows from detecting value differences or similarities between datasets. *Quality assessment* using among other *provenance* will be implemented too: What was done with the data, and how. This should be transparent for the user.

Metadata

The platform offers a metadata system, which is also linked to open data in order to have advanced search facilities. The metadata system is also a tool to support data integration, due to the fact that the dataset owner is stimulated to use URLs in the metadata (figure 6). And it is supported by the category services discussed above. (For more details: Idrissou et al (2015).

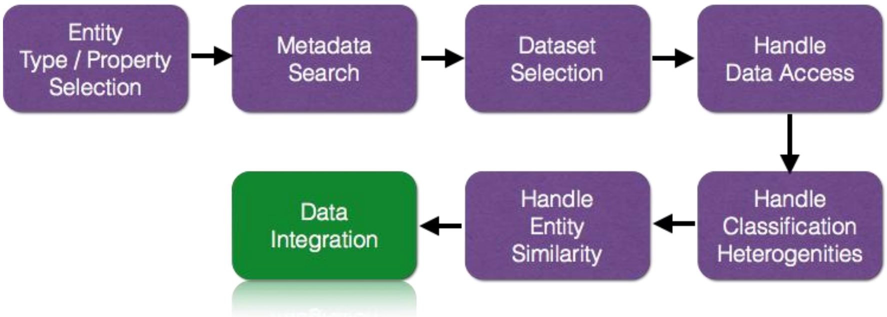
Figure 6: The RISIS/SMS metadata system



Access control
The platform provides access to a variety of datasets, of which some are open, some are proprietary and require e.g., subscription, and other are confidential. As data are only partly open, access control is provided and essential. With the type of data we are focussing on, privacy issues and legal issues may easily come up.

The workflow
From the users" perspective, the platform does two things. Firstly there is the workflow to identify data needed by the user to do a research project. This goes from identifying the entities and the variables (properties) needed. Through the metadata search the relevant datasets can be selected. If access can be given, steps follow like classification matching and disambiguation, and then the data can be integrated. The workflow is represented in figure 7.

Figure 7: The users workflow

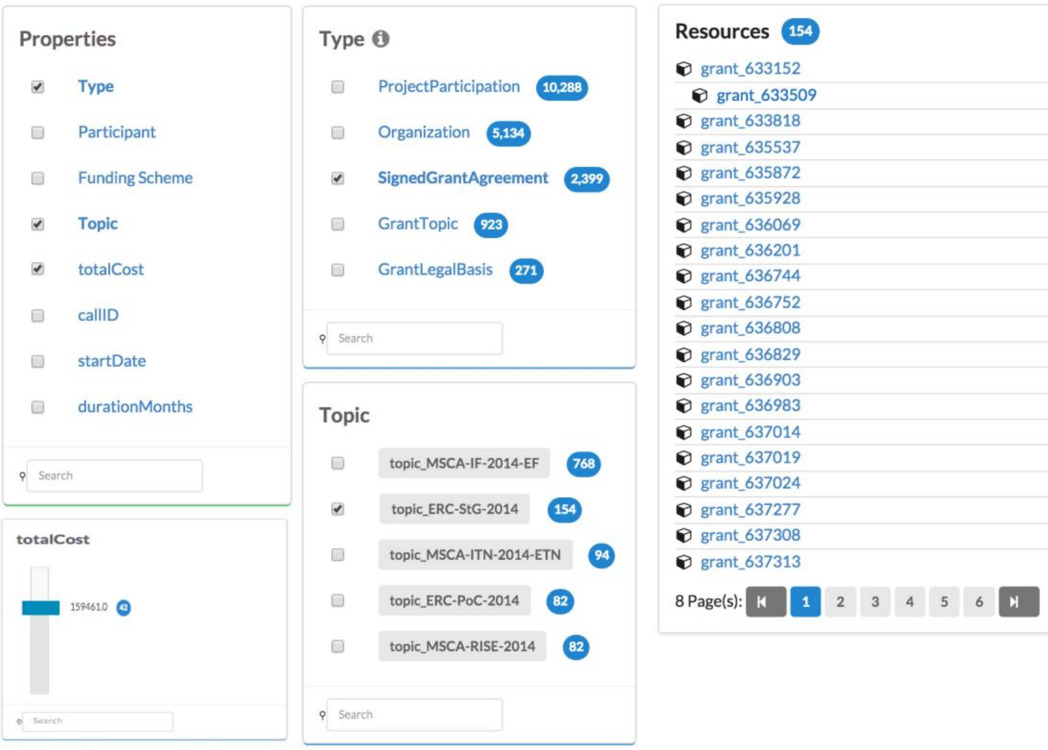


Secondly, when the integrated data are available, the user wants to have a dataset to do the analysis and visualization. (Standard) queries are provided to get the required data into the required format. Tis sounds simpler than it is, but experience with other data integration platforms show that the user needs support by specialists to query the platform. This suggests that it is indeed more an infrastructure than a tool. The output can have various formats, to enable deployment of general or specific analytical tools. A specific interface will be developed to connect the SMS platform to the Cortext platform (www.cortext.fr).

Data can also be browsed, in order to get a more qualitative feeling for the data. The *facet browser* is used for this (figure 8). Faceted browsing is particularly useful when you would like to present users with multiple entry points into a dataset or when there is no expectation that they know what they are looking for beforehand. It allows users to explore the space of potential items by choosing the refinements in any order.

Another use of the facet browser is when searching for information for more qualitative studies. The linked nature of the data enable to search for rich information about the entities one is interested in.

Figure 8. An example of SMS faceted browser



THE DEMONSTRATOR

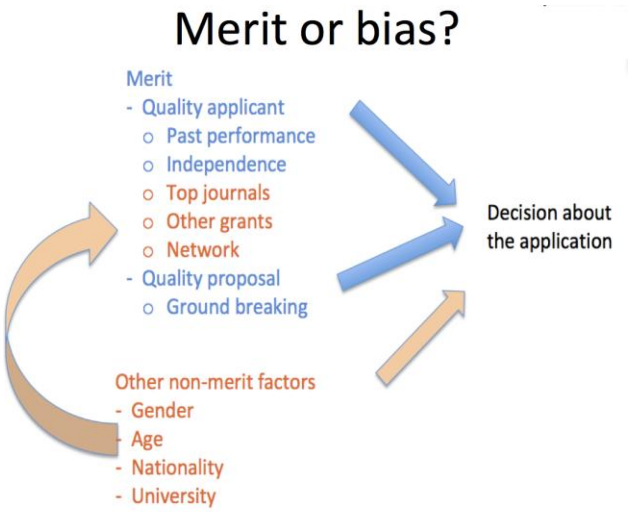
Many parts of the platform are already implemented and tested. Currently we are finalizing the beta-version of the platform, and the planning is that after this summer, the platform is available for the first users at http://sms.risis.eu.

We show the use of SMS in a demonstrator project, investigating gender bias in grant allocation. In this project we try to find out whether gender of applicants influences the grant decision. In order to answer that question, one needs to bring in a multitude of variables that may influence the decision – apart from gender. This may be variables representing merit – such as measures of scholarly performance, but also variable that measure performance in a possibly gendered way, such as the collaboration network. And it needs to include personal characteristics that can influence the decision, such as age, nationality, and so on. The model we use (figure 9) includes quite some – theory driven – variables. These variables come from a variety of data sources:

- From Web of Science: Bibliometric performance scores
- Quality of the applicants network: Organizations mentioned in the CV (PDF), and ranking of those organizations from Leiden Ranking (Excel)
- Earlier grants: from CV (PDF)
- Host institution from admin file (Excel) and ranking of host institution from administrative file (Excel)
- Personal characteristics from admin file (Excel)
- Linguistic categories in evaluation: Term extraction from review forms (PDF)

We used the SMS platform for pre-processing, for converting into RDF, for entity recognition and linking. The output is a data file for analysis.

Figure 9: the GendERC project – analytical model



Preliminary findings suggest that gender bias indeed exists, but different in the different disciplines. But for this paper, it is more interesting how it was done then what comes out (Van den Besselaar et al 2016; Van den Besselaar 2016)

CONCLUSIONS

We expect that platforms like SMS will enable research within the STI field that was not possible before. Studies can become large-scale, can including many more variables than traditionally has been the case. More and more appropriate data can be exploited. Within the (life) sciences, instrumentalities and infrastructures have radically changed the way research is done (de Solla Price 1984). In the social sciences and humanities this has been much less that case; but that may change in the near future. New data integration and enriching infrastructures may open the space of new forms of social science. As Nicholas Christakis (2013) wrote: "Let‟s shake up the social sciences".

REFERENCES

Beek W, Rietveld L, Schlobach S, Van Harmelen F, LOD Laundromat; Why the semantic web needs centralization (even if we don‟t like it). IEEE Internet Computing, March-April 2016

Christakis N, Let‟s shake up the social sciences" *New York Times* (July 19, 2013),

Ciccarese P, S. Soiland-Reyes, K. Belhajjame, A. J. Gray, C. Goble, and T. Clark. Pav ontology: provenance, authoring and versioning. *Journal of biomedical semantics*, 4 (2013) 1,1-22,

Daraio C, M. Lenzerini, C. Leporelli, H. F. Moed, P. Naggar, A. Bonaccorsi, and A. Bartolucci. Data integration for research and innovation policy: an ontology-based data management approach. *Scientometrics*, 1-15, 2015.

de Solla Price D, The science/technology relationship, the craft of experimental science and policy for the improvement of high technology innovation. *Research Policy* **13** (1984) 3-20

Groth, P, A. Loizou, A. J. Gray, C. Goble, L. Harland, and S. Pettifer. Api-centric linked data integration: The open fPHACTSg discovery platform case study. *Web Semantics: Science, Services and Agents on the World Wide Web* **29** (2014) 12-18

Gurney T, Horlings E, van den Besselaar P, Author Disambiguation Using Multi-Aspect Similarity Indicators. *Scientometrics* **91** (2012) 435-449

Idrissou, AK, Khalili A, Hoekstra R, van den Besselaar P, Managing metadata relevant for research and innovation studies: The RISIS case. Paper *Whise workshop* 2016

Khalili A, A. Loizou, and F. van Harmelen. Adaptive linked data-driven web components: Building fexible and reusable semantic web interfaces. *Extended Semantic Web Conference* (ESWC) 2016.

Sandström, U., & Sandström, E., The field factor: Towards a metric for academic institutions. *Research Evaluation* **18** (2009) 243–250.

Van den Besselaar, P, *Predicting panel scores by linguistic analysis*. Paper submitted to the STI conference, Valencia 2016

Van den Besselaar & Sandström, *What is the required level of data cleaning? A research evaluation case*. forthcoming

Van den Besselaar P, Schiffbanker H, Snaström U, Holzinger F, Polo L, *Explaining gender bias in grant selection – the ERC starting grants case*. Paper submitted to the conference on Gender and Higher Education, Paris 2016

Data Citation Policies of Data Providers within the scope of Longitudinal Studies in Life Course Research[1]

Anke Reinhold*, Marc Rittberger** and Nadine Mahrholz***

*reinhold@dipf.de
German Institute for International Educational Research (DIPF), Schloßstraße 29, Frankfurt, 60486 (Germany)

** rittberger@dipf.de
German Institute for International Educational Research (DIPF), Schloßstraße 29, Frankfurt, 60486 (Germany)

*** nadine.mahrholz@uni-hildesheim.de
University of Hildesheim, Universitätsplatz 1, Hildesheim, 31141 (Germany)

ABSTRACT
In this article, a small-scale case study analyzing the nature of data citation policies within the scope of longitudinal studies in life course research is presented. The sample consists of eight data providers from Europe, North-America and Australia and was evaluated with regard to eight criteria which potentially affect data citation behavior of researchers in the field, for example the wording of data citation obligations or sanctions for not citing research data in accordance to given requirements. The study demonstrates that research data providers follow a wide range of approaches to data citation, especially in terms of data citation location within a publication as well as disposal obligations for data-related publications. However, this diversity might lead to inconsistency in data citation behaviour and also to a general lack of comparability of data citation quantity and quality as relevant factors in research evaluation.

INTRODUCTION
In order to meet the requirements of funding organisations or policy makers, the scientific output of researchers, research groups, institutions and even countries is regularly tracked by indicators that measure for example citation rates or citation impact. With the rise of altmetrics, attention in research monitoring has also shifted towards research activities that are – exclusively or complementarily – visible on the social web. However, citation analysis is still mainly focused on publication-related research output and so far only a few works have discussed the distinctiveness of research data as a considerable factor in citation analysis and research evaluation. For example, quantitative analyses of the Data Citation Index (DCI) (Thomas Reuters) (e.g. Peters, Kraker, Lex, Gumpenberger & Gorraiz, 2015; Robinson-García, Jiménez-Contreras & Torres-Salinas, 2015) as well as subject-specific publication depositories (Mooney, 2011; Mooney & Newton, 2012) have shown a general uncitedness of research data in the social sciences and the humanities, despite the fact that sharing research data can be associated with higher citation rates (Piwowar, Day & Fridsma, 2007).

Studies analysing the quality of data citation behaviour also uncovered that data citation is not carried out adequately with regard to existing requirements of academic journals (Mooney & Newton, 2012) or research data providers (Mahrholz, Reinhold & Rittberger, 2015). Additionally, as argued by Robinson-García et al. (2015), the citedness of research data

[1] This work was supported by the Leibniz Institute for Educational Trajectories (LIfBi).

heavily depends on the quality of data-related information provided by data repositories and varies across disciplines. Furthermore, data citation policies of scientific journals tend to be slightly stricter in the natural sciences than in the social sciences (cf. Blahous et al., 2015). A case study analysing data citation and sharing policies in the environmental sciences also demonstrates that "an overwhelming majority of funding agencies, repositories and journals fail to provide explicit directions for sharing and citing data" (Weber et al., 2011, p. 1). Obviously, making research data accessible and usable is a time-consuming and cost-intensive task. As a consequence, these activities should be appreciated by the scientific community and moreover demand for the inclusion of data citation indicators as a relevant factor in research monitoring. However, in order to make valid statements about data citation quantity and quality, it is necessary to thoroughly analyse the nature of data citation policies within a certain domain. In this paper, data citation policies of eight research data providers in Europe, the United States and Australia within the scope of longitudinal studies in life course research are being evaluated, e.g. with regard to citation principles and sanctions for data users who do not cite adequately. The aim of the study is to outline the different approaches followed by data providers or data repositories in terms of data citation policies which might influence data usage and citation behaviour of researchers in the domain.

Life course research is currently a very dynamic field of research in the social sciences. It provides stakeholders in politics and education with extensive and reliable data about life paths, transitions and decisions in private as well as professional lives. Furthermore, societal changes over extended timeframes of several years or even decades are being monitored. Longitudinal studies in live course research are generally characterized by large sample sizes, different cohorts of participants and various waves of surveys. There is also a strong demand for protecting sensitive personal information, e.g. about performance in school or the parent-child relationship, which are retrieved in these studies at a large scale. As a result, data providers in life course research generally dispose of high data security standards and offer a variety of data access modes, different type of data formats and data granularity. Users generally have to commit to data use agreements and are obliged to cite the research data used according to specific requirements. These data citation policies include aspects of contractual obligations of data citation, concrete requirements of including data citation elements (e.g. a persistent identifier) (cf. Mooney & Newton, 2012) or the position of the data citation within a publication (e.g. in the abstract or the references section) as well as disposal obligations for publications based on the research data provided.

DATA CITATION POLICIES OF DATA PROVIDERS IN LIFE COURSE RESEARCH – A CASE STUDY

For the case study a sample of eight longitudinal studies across the life course in Europe, North-America and Australia was identified by means of six criteria to ensure comparability: 1) thematic focus on educational and personal transitions, 2) ongoing research project, 3) at least a national or international perspective, 4) elaborated data access technologies (e.g. via a data center), 5) data use agreements as a prerequisite for data usage of sensitive data, 6) mention of data citation requirements.[2] Based on these criteria the following longitudinal studies were selected:

[2] The criteria were applied to the result set of an extensive web search which retrieved overall 19 longitudinal studies across the life course in Europe, North-America and Australia. The starting point for the web search was a list of longitudinal studies in the social sciences issued by Mallock, Riege & Stahl (2016, p. 146-148).

Table 1. Sample of longitudinal studies across the life course.

Study name	**Research topics**	**Country**	**Start in year**
Étude Longitudinale Française depuis l'Enfance (ELFE)	Impact of family circumstances, living conditions and environment on the physical and psychological development, health and socialization of children.	France	2011
Millennium Cohort Study (MCS)	Influence of early family context on child development and outcomes throughout childhood, adolescence and adulthood.	UK	2000
Negotiating the Life Course	Changing life courses and decision-making processes of men and women as the family and society move from male breadwinner orientation in the direction of higher levels of gender equity.	Australia	1997
National Educational Panel Study (NEPS)	Educational processes from early childhood to late adulthood.	Germany	2009
Panel Analysis of Intimate Relationships and Family Dynamics (pairfam)	Partnership and family dynamics in Germany.	Germany	2008
Socio-Economic Panel (SOEP)	Objective living conditions, values, willingness to take risks, current social changes, and the relationships and interdependencies among these areas.	Germany	1984
Transitions from Education to Employment (TREE)	Post-compulsory educational and labour market pathways of school leavers.	Switzerland	2001
Panel Study of Income Dynamics (PSID)	Employment, income, wealth, expenditures, health, marriage, childbearing, child development, philanthropy, education, and numerous other topics.	US	1968

For each of the research data providers in the domain of longitudinal studies across the life course, the following eight factors were documented by thoroughly eliciting regulatory and user service information on the data providers web sites[3]: 1) *wording of obligations with regard to data citation*, 2) *requirements for obligatory data citation elements*, 3) *requirements for data citation location within a publication*, 4) *availability of concrete examples for data citation*, *5) obligation to report data-related publications*, 6) *period of notification for data-related publications,* 7) *disposal obligation*[4] *for data-related publications* and 8) *sanctions for*

[3] For the analysis, different information sources on the providers' websites were reviewed, e.g. the data use agreements or the specific data citation section. The URLs of the homepages of all data providers in the sample are mentioned in the reference section.
[4] Publications which are based on a specific dataset are to be submitted to the data provider as a paper-based or digital version according to an agreement of use.

not citing research data in accordance to the requirements. From the point of view that the citedness of research data heavily depends on the quality of data-related information provided by data repositories (cf. Robinson-García et al., 2015), it is legitimate to assume that all of these factors might affect data citation behaviour of researchers in the field.

FINDINGS AND DISCUSSION

All eight data providers issue data use agreements that oblige their users to cite research data. The wording of these obligations (1) in the data use agreements differs significantly, ranging from very concrete citation specifications to rather general requests to cite in accordance to "academic conventions". Furthermore, all providers name obligatory data citation elements (2): Seventy-five percent of the data providers in the sample demand for including a distinct data version, 50% for including a Digital Object Identifier (DOI)[5] and 37.5% for naming a specific reference article which outlines the original study design. All eight data providers ask for the inclusion of an acknowledgement phrase indicating either the name of the study or the data center involved. These findings clearly indicate that data providers in life course research generally follow a top-down approach to prevent uncitedness of research data. It is also noticeable that again only 37.5% of providers in the sample provide guidelines for data citation location within a publication (3), e.g. for citing the study as the originator of the data in the title, the abstract or the reference section. This is surprising as it can be assumed that these recommendations are not only useful for guiding data users in the writing process. The recommendations might also foster awareness amongst researchers about the "quality" of a data citation within a document. For example, a data citation in the title or in the abstract can possibly be assessed as more valuable than a data citation in the caption of table or a figure. Interestingly, the data use agreement of the French ELFE study already indicates that users are obliged to cite the study in the title *and* the body of the text if the article is exclusively or primarily based on ELFE data (ELFE, 2014).

Apart from one, all data providers publish concrete examples for data citation on their websites which for example include the names of the authors (of a reference article), the name of the study and the DOI (4). Of course, researchers can already refer to more general data citation guidelines (cf. DataCite, 2014; ESRC, 2016; ZBW, GESIS & RatSWD, 2015[6]). Precise citation examples which relate to the actual study in use might nevertheless be even more important for supporting researchers and help them to prevent citation errors. Seventy-five percent of the providers in the sample insist on the obligation to report data-related publications (5) with only one provider, the Leibniz Institute for Educational Trajectories (LIfBi) for the NEPS data, calling for a period of notification for data-related publications of four weeks before publishing (6). And 50% of the data providers even issue a disposal obligation for publications using research data (7). Surprisingly, only one data provider – again LIfBi – calls for sanctions if data users do not cite in accordance to the data use agreement (cf. LIfBi, 2015) (8)[7]. In summary, it might be assumed that research data providers have already identified a need for action with regard to data citation misbehaviour. However, it still needs to be verified whether the data citation policies described here are

[5] One data provider has just recently added the obligatory inclusion of a DOI in his citation recommendations – this might be an indicator that the DOI becomes more widely accepted within the domain.
[6] This publication is not available in English yet.
[7] Although it is not explicitly stated that non-citations cause a breach of contract, citing the study name and the dataset used for analysis can be interpreted as "essential obligations" of the data use agreement.

appropriate measures for achieving high citation rates and citation quality of research data issued by providers of longitudinal data in life course research.

As stated in the introduction, the main goal of the study was to outline the variety of data citation policies within life course research and to discuss possible implications for data use and citation behaviour in the field. It could be demonstrated that data providers follow differing approaches in terms of data citation requirements. This involves data versions, identifiers and reference articles describing the original study design. In addition, data providers differ substantially with regard to recommendations for data citation location as well as disposal obligations for data-related publications. This might lead to a high diversity in data citation behaviour of data users in the field and potentially to non-comparable results in data citation analysis. It is therefore reasonable to argue that data providers should pursue the harmonisation of data citation specifications – in close cooperation with journals and research institutions involved in life course research. Furthermore, policy makers should strongly encourage the development of domain-specific data citation indicator sets for the valid representation of scientific output, allowing for an improved comparability and traceability of research.

LIMITATIONS OF THE STUDY AND OUTLOOK

We are aware that our research has some limitations. First, the study consists of a small sample which is not representative for data usage and citation within the social sciences in general. Second, there might be other longitudinal studies in life course research that meet the selected criteria presented above. Third, there is a predominance of European longitudinal studies in the sample. Finally, the study does not investigate the influence of data citation policies on the actual data citation behaviour of researchers in the field. A consecutive study, analysing data citation quantity and quality in a large sample of data-related publications in the social sciences might substantially enhance our understanding of data citation behaviour. Despite these limitations we believe our work has highlighted the importance of critically examining data citation policies beforehand as one milestone of coherent and comparable data citation analysis.

REFERENCES

Blahous, B., Gorraiz, J., Gumpenberger, C., Lehner, O., Stein, B. & Ulrych, U. (2015). Forschungsdatenpolicies in wissenschaftlichen Zeitschriften – Eine empirische Untersuchung. In *ZfBB*, 62, 12-24.

DataCite International Data Citation Metadata Working Group (2014). DataCite Metadata Schema for the Publication and Citation of Research Data. Version 3.1 October 2014. Retrieved March 16, 2016 from: http://schema.datacite.org/meta/kernel-3/doc/DataCite-MetadataKernel_v3.1.pdf.

Étude Longitudinale Française depuis l'Enfance (ELFE) (2014): Charte d'accès aux données Elfe. Retrieved March 16, 2016 from: http://www.elfe-france.fr/images/documents/charte_acces_donnees_plateforme.pdf.

Economic and Social Research Council (ESRC) (2016). Data Citation. What you need to know. Retrieved March 16, 2016 from: https://www.ukdataservice.ac.uk/media/104397/data_citation_online.pdf.

Leibniz-Institute of Educational Trajectories (LIfBi) (2015). NEPS. Datennutzungsvertrag. Retrieved March 16, 2016 from: https://www.neps-data.de/Portals/0/NEPS/Datenzentrum/Datenzugangswege/Vertraege/NEPS_Datennutzungsvertrag_DE.pdf.

Mahrholz, N., Reinhold, A. & Rittberger, M. (2015). Data Citation Quantity and Quality in Research Output of a Large Educational Panel Study. In *Proceedings of the 15th International Conference of Knowledge Technologies and Data-Driven Business* (I-Know '15). Graz: ACM.

Mallock, W., Riege, U., Stahl, M. (2016). *Informationsressourcen für die Sozialwissenschaften. Datenbanken, Längsschnittuntersuchungen, Portale, Institutionen*. Wiesbaden: Springer.

Millennium Cohort Study (MCS). Retrieved March 16, 2016 from: http://www.cls.ioe.ac.uk/page.aspx?&sitesectionid=851&sitesectiontitle=Welcome+to+the+Millennium+Cohort+Study.

Mooney, H. and Newton, M. P. (2012). The Anatomy of a Data Citation: Discovery, Reuse, and Credit. In *Journal of Librarianship and Scholarly Communication*, 1, 1, eP1035.

Mooney, H. (2011). Citing data sources in the social sciences: do authors do it? In *Learned Publishing*, 24, 2, 99-108.

Negotiating the Life Course. Retrieved March 16, 2016 from: http://lifecourse.anu.edu.au/.

Panel Analysis of Intimate Relationships and Family Dynamics (pairfam). Retrieved March 16, 2016 from: http://www.pairfam.de/en/.

Panel Study of Income Dynamics (PSID). Retrieved March 16, 2016 from: https://psidonline.isr.umich.edu/.

Peters, I., Kraker, P., Lex, E., Gumpenberger, C. & Gorraiz, J. (2015). Research Data Explored: Citations versus Altmetrics. In *Proceedings of the 15th International Conference on Scientometrics and Informetrics* (ISSI '15) (p. 172-183). Istanbul: Bogaziçi University Printhouse.

Piwowar, H.A., Carlson, J.D. & Vision, T.J. (2011). Beginning to track 1000 datasets from public repositories into the published literature. In *Proceedings of the American Society for Information Science and Technology*, 48, 1, 1-4.

Piwowar, H.A., Day, R.S. & Fridsma, D.B. (2007). Sharing Detailed Research Data Is Associated with Increased Citation Rate, *PLoS ONE* 2, 3, e308. Retrieved March 16, 2016 from: http://www.ncbi.nlm.nih.gov/pmc/articles/ PMC1817752/.

Robinson-García, N., Jiménez-Contreras, E. & Torres-Salinas, D. (2015). Analyzing data citation practices using the data citation index. In *Journal of the Association for Information Science and Technology*, DOI: http://dx.doi.org/10.1002/asi.23529 (Preprint).

Socio-Economic Panel (SOEP). Retrieved March 16, 2016 from: http://www.diw.de/en/soep.

Transitions from Education to Employment (TREE). Retrieved March 16, 2016 from http://www.tree.unibe.ch/index_eng.html.

Weber, N.M., Piwowar, H.A. & Vision, T.J. (2010). Evaluating Data Citation and Sharing Policies in the Environmental Sciences. In *Proceedings of the American Society for Information Science and Technology*, 47, 1, 1-2.

ZBW, GESIS & RatSWD (2015). Auffinden, Zitieren, Dokumentieren: Forschungsdaten in den Sozial- und Wirtschaftswissenschaften. Retrieved 16 March, 2016 from: http://auffinden-zitieren-dokumentieren.de/wpcontent/uploads/2015/03/Forschungsdaten_DINA4_ONLINE_VER_02_06.pdf.

Stepping up Information Infrastructures and Statistical Reporting – Monitoring the German Excellence Initiative

Anke Reinhardt

*anke.reinhardt@dfg.de
Information Management Unit, Deutsche Forschungsgemeinschaft (DFG, German Research Foundation), Kennedyallee 40, 53175 Bonn (Germany)

ABSTRACT
The Excellence Initiative has not only been the most prominent funding scheme in German research policy in recent years, but has also had important side effects on research management. This paper argues that the Excellence Initiative was indeed a "boost" for improving the data infrastructure and statistical reporting of the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation). The learning effects are now transferred to the line business and serve as a good starting point for the reporting on a potential third phase of the Excellence Initiative.

BACKGROUND
The Excellence Initiative is a funding scheme launched in 2005 with the aim of promoting top-level research in Germany. It has three funding lines: "Institutional Strategies for Top-Level University Research" are supposed to increase the international competitive ability of the entire university; "Graduate Schools" (GSC) should provide highest-level research training; "Clusters of Excellence" (EXC) pool excellent researchers in interdisciplinary centres.

The DFG is the largest funding agency for basic research in Europe, supporting almost 30 000 research projects in all scientific disciplines (Deutsche Forschungsgemeinschaft 2015a). Bund and Länder jointly commissioned the DFG – together with the Wissenschaftsrat (WR, German Council of Science and the Humanities) – to implement the funding scheme on their behalf, and with an additional budget.

So far, the Excellence Initiative runs until 2017 for two funding periods (2006-2012 and 2012-2017). After DFG and WR provided a report about the implementation of the Excellence Initiative in 2008 (Sondermann et al. 2008), Bund and Länder granted a second round of funding. However, they also agreed that a decision about a renewal of the Excellence Initiative after 2017 should be based on an external evaluation ("Imboden Commission"). To support this evaluation, in 2015 DFG and German Council of Science and Humanities (WR) were again expected to report on the course of the Excellence Initiative, this time in a "data-based way".

the challenge of the "data-based report"
The purpose of the data-based report was to provide information on effects and – if possible – output of the funded projects. Of particular interest were structural effects, e.g. the working conditions of PhD candidates, international appeal and recruitment successes of universities, interdisciplinarity and cooperation ties with non-university research institutes or industry.

The usual source for statistical services of the DFG is its internal database, which stores process-produced data on 620,000 applications (e.g. requested resources, discipline, time span, collaborators), on applicants and reviewers (gender, age, nationality etc., in total 230,000 scientists) and on 50,000 research institutes. Besides the actual grant management, the DFG uses this data to steer the programmes and to perform detailed analyses on all aspects of its funding, e.g. success-factors (Dinkel & Wagner 2015), participation of women (DFG 2015b), internationalization (Fuß 2011) or interdisciplinarity (Güdler 2013). For in-depth analysis, it commissions evaluation studies, which also include additional data sources, e.g. interviews, bibliometrics, surveys etc. (Reinhardt 2013).

The DFG, being responsible for the data-based reporting on EXC and GSC, had to focus much more on throughput and output information than it usually does. It decided to include other data sources in addition to its internal database. Additionally, it needed a different analytical perspective on the effects of science funding, with the particular difficulty that almost all universities in Germany participated in the Excellence Initiative.

NOVELTIES IN THE DATA INFRASTRUCTURE AND LINES OF ANALYSIS

To develop the concept of the data-based report, the DFG established a new organisational setting: It consulted with a scientific advisory board of five eminent scientists in the field of research evaluation and science of science studies on the question which data to collect and how to analyse it.

One measure was that the DFG started to collect data on the "life" in projects. EXC and GSC have a large number of scientific members (usually between 100 and 600) not involved in the application for funding, who are therefore "unknown" to the DFG data-base. The DFG collected information on the doctoral candidates, the postdocs, research group leaders, guest researchers and other scientific staff, together more than 20,000 persons, as well as on professorial positions created.

Additionally the DFG hired a contractor to analyse renewal applications of GSC and EXC and of the protocols of the peer review panels to identify oganisational measures and structural effects mentioned there. An online survey of 990 involved researchers as well as interviews and focus groups, mainly with Principal Investigators and presidents of universities, asked about experiences and opinions of the Excellence Initiative. A survey of reviewers complemented the picture (Möller 2012).

The report to Bund and Länder summarised information from all these sources. Additionally to this report, the DFG published analyses that dig deeper into specific effects.

One of the most prominent reporting products of the DFG is the Funding Atlas (DFG 2015c). It provides information on public research funding in Germany, particularly DFG funding, at German universities. The 2015 edition specifically looked at concentration effects of the Excellence Initiative on funding and on disciplinary profiles of universities using the Gini coefficient. It turns out that the Excellence Initiative did not increase the concentration but instead leveraged more grant-seeking activity at all German universities. It strengthened subjects that were strong before. Network analysis of disciplines, based on the classification of proposals, and on regions, based on researcher's location, deepened the understanding of the German research landscape. Additionally a bibliometric analysis of two subjects indicates that the funded universities were slightly more productive than others [1].

DISCUSSION

The political importance attached to the Excellence Initiative proved to be a catalyst for change. The DFG used some of the novelties introduced in the Excellence Initiative to step up its data infrastructure and reporting more generally. For example, it transformed the survey instrument introduced in the Excellence Initiative to also survey the CRC and RTG, which in the future allows to compare these funding lines. Equally, it uses the document analysis methodology used in the Excellence Initiative in a project analysing the effects of its "Research Oriented Standards on Gender Equality".

After the Excellence Initiative is before the Excellence Initiative: Currently politics debates about the shape of a future round. While it has already agreed on an extension, the exact format will only be decided in June 2016. However, it seems that in the future even more focus will lay upon the effects and side-effects: Will the large number of PhDs educated in the GSC be able to find qualified jobs? Are the recruited researchers there to stay? Has the governance of universities changed for good? Answering these questions requires other kinds of data and other approaches than the ones used previously.

The DFG will therefore place even more emphasis on output data. Final reports are a good source of information on publications, scientific content and staff. The DFG plans to start collecting final reports electronically which allows to analyse the information more easily. A specific focus will be on the text analysis of proposals, on tracking research topics, and on career outcomes by researching the placement of staff members.

In the meantime, new data sources are available. For example, funding acknowledgements allow links between funding and specific publications as well as their citation rates. The DFG needs to enforce its policy on this. A new statistics on doctoral researchers by the German Federal Statistical Office can supplement information that the DFG collects in its survey.

The more universities are involved in the Excellence Initiative, the harder it will become to single out the effects of the Excellence Initiative. The DFG is therefore eager to cooperate again with researchers to ask the rights questions about this funding instrument and to test novel methodologies of analysis. This will allow gaining a deeper understanding not only on the Excellence Initiative but on research funding more general.

[1] The DFG was not alone in analysing the effects of the Excellence Initiative. For example, the Berlin-Brandenburg Academy of Sciences and the Humanities published a bibliometric report on the relative success of universities funded by the Excellence Initiative versus others in terms of publications and citations (Hornbostel & Möller 2015). Engels et al. analysed its effects on Gender equality (Engels et al. 2015).

REFERENCES

Deutsche Forschungsgemeinschaft (2015a): Jahresbericht 2014, Bonn: DFG. www.dfg.de/dfg_profil/jahresbericht

Deutsche Forschungsgemeinschaft (2015b): Chancengleichheits-Monitoring 2014. Bonn: DFG. www.dfg.de/en/research_funding/principles_dfg_funding/equal_opportunities/monitoring_equal_opportunity

Deutsche Forschungsgemeinschaft (2015c): Förderatlas 2015. Kennzahlen zur öffentlich finanzierten Forschung in Deutschland. Bonn: DFG.

Dinkel, W. & J. Wagner (2015): Who applies for DFG grants? Statistical Trends in Proposal Submission and Applicant Demographics. *DFG Infobrief*, 1. Bonn: DFG.

www.dfg.de/en/dfg_profile/facts_figures/evaluation_studies_monitoring/info_briefe

Engels, A. et al. (2015): Bestenauswahl und Ungleichheit. Eine soziologische Studie zu Wissenschaftlerinnen und Wissenschaftlern in der Exzellenzinitiative. Frankfurt a.M.: Campus.

Fuß, K. (2011): Doctoral Training Frameworks. A Comparison of Doctorates in the DFG's Research Training Groups and Collaborative Research Centres Programmes. *DFG Infobrief*, 4. Bonn: DFG.

www.dfg.de/en/dfg_profile/facts_figures/evaluation_studies_monitoring/info_briefe

Güdler, J. (2013): Interdisciplinary review processes: structural impact and funding success. Bonn: DFG.

www.dfg.de/dfg_profil/zahlen_fakten/evaluation_studien_monitoring/studien/bericht_interdisziplinaritaet/

Hornbostel, S., & Möller, T. (2015): Die Exzellenzinitiative und das deutsche Wissenschaftssystem: Eine bibliometrische Wirkungsanalyse. *Wissenschaftspolitik im Dialog 12*. Berlin: Berlin-Brandenburgische Akademie der Wissenschaften.

Möller, T. et al. (2012): Exzellenz begutachtet. Befragung der Gutachter in der Exzellenzinitiative. *iFQ-Working Paper, 11*. Berlin: iFQ.

Reinhardt, A. (2013): Das Evaluationsvorgehen der Deutschen Forschungsgemeinschaft. *Zeitschrift für Evaluation*, 2, pp. 305-314.

Sondermann, M. et al. (2008): Die Exzellenzinitiative. Beobachtungen aus der Implementierungsphase. *iFQ-Working Paper*, 5, Bonn: iFQ.

CHAPTER 2

Smart Use of Indicators for Innovation Policy

Innovation indicators: Towards a User's guide

Michiko Iizuka* and Hugo Hollanders**
*iizuka@merit.unu.edu
UNU-MERIT, Boschstraat 24, Maastricht, 6211AX (The Netherlands)

** h.hollanders@maastrichtuniversity.nl
UNU-MERIT, Boschstraat 24, Maastricht, 6211AX (The Netherlands)

ABSTRACT
The increased use of innovation indicators is observed in policy sphere. Several reasons are attributed for this increased use: first, access to data on innovation increased as the result of over half a century of efforts made by international organizations and researchers in this field; second, an increase in types and number of innovation indicators; third, its increased use in business contexts with the understanding that innovation is an integral part of business performance, which further spurred the use and development of innovation indicators (Soete and Freeman, 2009); fourth, the expansion in geographical coverage of countries, particularly in emerging countries of Latin America, Asia and Africa (Gault, 2010) making innovation indicators more policy (politics) relevant in a wider variety of countries, e.g. by allowing to benchmark and compare countries' innovation performance; and fifth, the innovation indicators (in particular, the composite indicator but any indicator if used in ranking countries) became a communicative tool in public debate, in the backdrop of recent emphasis on 'evidence based policy' and 'participatory decision making' in the policy domain (OECD, 2012).

Considering innovation indicators are intended to improve the performance of innovation policy, their increasing use is generally good news. Nevertheless, due to the increase in diversity in type and context in which indicators are being applied there is an increased occurrence of inappropriate use and misinterpretation of innovation indicators in the policy sphere. This paper first describes the specific cases then tries to generalize the problem aiming to build a general guideline or check list on the appropriate use of innovation indicators.

1. MOTIVATION/INTRODUCTION

The increased use of innovation indicators is observed in policy sphere. Several reasons are attributed for this increased use: first, access to data on innovation increased as the result of over half a century of efforts made by international organizations and researchers in this field; second, an increase in types and number of innovation indicators; third, its increased use in business contexts with the understanding that innovation is an integral part of business performance, which further spurred the use and development of innovation indicators (Soete and Freeman, 2009); fourth, the expansion in geographical coverage of countries, particularly in emerging countries of Latin America, Asia and Africa (Gault, 2010) making innovation indicators more policy (politics) relevant in a wider variety of countries, e.g. by allowing to benchmark and compare countries' innovation performance; and fifth, the innovation indicators (in particular, the composite indicator but any indicator if used in ranking countries) became a communicative tool in public debate, in the backdrop of recent emphasis on 'evidence based policy' and 'participatory decision making' in the policy domain (OECD, 2012).

Considering innovation indicators are intended to improve the performance of innovation policy, their increasing use is generally good news. Nevertheless, due to the increase in diversity in type and context in which indicators are being applied there is an increased occurrence of inappropriate use and misinterpretation of innovation indicators in the policy sphere. This paper first describes the specific cases then tries to generalize the problem aiming to build a general guideline or check list on the appropriate use of innovation indicators.

2. PROBLEMATIC USE OF INNOVATION INDICATORS IN THE POLICY DOMAIN

Some of the problematic use of innovation indicators is categorized into five subheadings. The first problematic use concerns the misinterpretation of innovation indicators due to a basic misconceived notion associated with innovation, such as R&D is a precondition for innovation, more or higher performance on a particular innovation indicator is always better and more innovation lead to positive outcomes. The second problem concerns the inappropriate use of innovation indicators. The two problems listed under this heading are similar in adapting a simplistic view of a complex reality: one problem concerns the construction or design of an indicator and the other problem the context in which it is applied. In anotherwords, an indicator is often used without a clear understanding of its construction (design) or even if the user understands the design of the indicator, the user does not know whether the existing design can be used in a different context (be it industry, sector or country). The third problematic use concerns the misuse of an innovation indicator in the policy domain. This problematic use is very much related to the two before-mentioned subheadings, misinterpretation and inappropriate use, but is more strongly linked to the use for policy purposes. The fourth problematic use concerns mainly the 'unknown' part of an innovation indicator due to the current 'omission' of some of the factors selected to be used as 'indicator'. The selection of relevant indicators to get a grasp of a complex and multi-dimensional reality such as innovation is a difficult task. Furthermore, as innovation is a dynamic process, some factors selected as indicators can become obsolete in the new context while other factors gain importance but it is simply 'not known' at the moment (an example would be the ownership of mobile phone or access to internet). The last problematic use concerns the issue of practices regarding the use of indicators in particular addressing the needs or use of indicators. Indicators are made to be used for improving innovation policy but sometimes there are mismatches in how indicators are designed and delivered and how indicators were demanded to be used and delivered.

In the following section, the problematic issues of innovation indicators in policy use are discussed by type of innovation indicators categorized as follows: (1) Traditional indicators used as innovation indicators (such as R&D expenditures, Patent applications, Human Resources and Scientific publications); (2) Innovation indicators collected through surveys; (3) Composite indicators.

BOX1 The 'problematic' use of innovation indicators in policy domain identified

1. Misinterpretation of innovation indicator

Assuming linear progression of innovation that R&D precedes innovation

Assuming that more/higher performance on an indicator is always better (Foray and Hollanders, 2015);

Assuming that more innovation automatically leads to development (Soete, 2013).

2. Inappropriate use of innovation indicator

Compare indicators, which are not comparable (due to different collection methods, assumptions, measurements, industrial structure);

Applying same indicator criteria in different sector and country setting without careful consideration of characteristics or new contexts in which the indicators are being applied.

3. Misuse of innovation indicator for policy purposes

Blindly applying the R&D/GDP target (in developing countries, it is 1% and in developed countries, it is 3%) as a policy goal without understanding a country's industrial structure and HR composition;

Applying the indicator to policy formulation without understanding the underlying conceptual design and data collection procedure;

Ignoring the country/sector/industrial structural context when interpreting the innovation indicators for policy purposes;

Relying only on composite indicator ranking to monitor, evaluate and formulate innovation policy (and to make political statements that would mislead the public).

4. Overlooked issues of innovation indicators in policy domain

Omitting important sources of innovation which are actually vital for the economy and therefore for policy formulation (in developed context, e.g. vocational education as an alternative to tertiary education (Foray and Hollanders, 2015) Globalization of business activities in developed context (Edquist and Zabala-Iturriagagoitia, 2015), Non R&D oriented innovation (i.e user led innovation/household innovation, public sector innovation?), in developing countries, informal R&D (user led innovation/household innovation), those who are trying to innovate from those who do not do anything, External sources of knowledge (Foreign Direct Investment (FDI), exports, GVC), informal sector, other productive sectors important in developing countries).

Ignoring the dynamic nature of industrial structure and relevance of selected indicators.

5. Mismatch of needs between user and producers of innovation indicator

a. Ignoring the results of innovation survey for policy elaboration (because the data come late (questions about its usefulness))

b. Ignoring the importance of comparability for the indicators (changing questions not to add).

Source: authors

3. THE PROPORTION OF RESEARCH AND DEVELOPMENTAL EXPERIMENTATION (R&D) IN GDP

R&D data are the most available data for the longest period of time and covers many countries. The concept and specification of R&D statistics are defined by the Frascati manual. Basically, R&D covers basic research, applied research and experimental development. This definition appears relatively straightforward but when data on R&D are to be collected it is not easy to distinguish research activities from non-research activities because the distinction between these two is determined by the 'intention' of actions taken. For example, 'action of taking temperature measurement' can be categorized as research and experimentation if the 'intention' was for research and experimentation while if the purpose was 'routine' activities, it is not counted as R&D activities.

Share of R&D in GDP is often used as an policy guideline to improve the innovation policy. For instance, targeting a certain percentage of R&D spending in GDP is often used as the policy goal. In fact, for the European Union, the Lisbon agenda sets 3% as the target; while many African and Latin American countries have 1% as their goal. While this can be useful as a general guideline, blindly applying the target to different country contexts, assuming that a higher percentage of GDP spent on R&D would lead to development misleads innovation policy. The reasons for such are as follows:

Appropriate to industrial structure

First, R&D intensities differ across industrial activities. Countries with different industrial structures should have different levels of the percentage of R&D that is appropriate for a given industrial structure. In another words, policy makers should pay attention to the efficiency and match of R&D expenditure to the needs of their country's industrial sector not just to increasing the share of R&D in GDP.

For instance, the OECD classifies the types of industries by the R&D intensities (e.g. in terms of R&D as a share of value added, R&D as a share of production, R&D plus technology embodied in intermediate and investment goods as a share of production). Currently the OECD uses a four-tier model to classify industries with R&D intensities are follows:

Box 2 R&D intensity

	Direct + indirect R&D as a share of production	R&D as a share of production	R&D as a share of value added
High tech industries	Above 7.5%	Above 7.5%	Above 15%
Medium high tech industries	Between 2.5% and 7.5%	Between 1.5% and 7.5%	Between 4% and 15%
Medium-low tech industries	Between 1% and 2.5%	Between 0.5% and 1.5%	Between 1.5% and 4%
Low tech industries	Below 1%	Below 0.5%	Below 1.5%

Source: Hatzichronoglou, 1997

Note: High tech industries include Aircraft and spacecraft, Pharmaceuticals, Office, accounting and computing machinery, Radio, TV and communications equipment, Medical, precision and optical instruments. Medium-high technology industries include Electrical machinery and apparatus, Motor vehicles, trailers and semi-trailers, Chemicals excluding pharmaceuticals, Railroad equipment and transport equipment, Machinery and equipment. Medium-low technology industries include Building and repairing of ships and boats, Rubber and plastics products, Coke, refined petroleum products and nuclear fuel, Other non-metallic mineral products, Basic metals and fabricated metal products. Low tech industries include Manufacturing, Recycling, Wood pulp paper, paper products printing and publishing, Food products, beverages and tobacco, Textiles, textile products, leather and footwear.

The categorization of technological level had slightly changed recently (2011) but overall principle of associating the type of activities (ISIC code) to the level R&D intensity continues (see technical note, OECD, 2011 and Hatzichronoglou, 1997).

Countries with higher shares of high tech industries are more likely to have higher share of R&D in GDP while countries with higher shares of medium-low tech industries (like Southern European countries) would have lower shares of R&D in GDP. In this way, appropriate level of the 'optimal' share of R&D in GDP can differ due to the industrial structure of a country. Hence a 3% or 1% guideline should be taken only as the guideline and not to be applied blindly in policy.

<u>Bias towards manufacturing/high tech sector</u>
Second, by design, R&D measurements are highly biased towards the manufacturing sector. This would create a problem when different sectors such as service, agriculture and natural resource based activities are to be assessed applying the same methods. This point is already being identified by the OECD. The technical notes of OECD directorate for STI states that "Direct R&D intensities are not much help for service activities. Instead other indicators such as skill intensity (e.g. education levels in industry x occupation matrices) and indirect R&D measures such as technology embodied in investment or investment in ICT goods by industry must be explored." (OECD, 2011). The same document also admits the limitation in disaggregating low tech industries due to the limited detailed R&D expenditure data across countries. On low tech industries, several studies also question the underlying assumption associated with low tech and low knowledge/technology intensity (Hirsch-Kreinsen and Schwinge, 2014, von Tunzelmann and Acha, 2005, Mendoca and von Tunzelmann, 2004). Hence, applying the preconceived notion from a particular context cannot be applicable to measure the conditions in different countries. In another words, the proportion of R&D in GDP cannot be used as the sign for the innovativeness of a country.

<u>Different policy implication due to the origin of R&D funding</u>
Third, differences in the origin of R&D (public versus private) are another distinction that needs to be considered. In general, developed countries have larger proportions of R&D performed and financed by the private sector while in developing countries the major contribution to R&D is made by the public sector. This difference will have different policy implications. In countries where the private sector is more active in R&D, policies targeting the private sector (policies such as tax incentives, subsidies etc.) can boost the share of R&D in GDP by raising business R&D expenditures.   If the share of R&D is larger in the public sector, then increasing the R&D would need to be preceded with policies to enhance human resources to carry out R&D and investment in the public research infrastructure (laboratories, university and research institutions, administrative capacities to carry out R&D).

More firms conduct innovation than R&D

One of the misinterpretations that is easily identified in developing countries is the assumption of linear progression that innovation always comes after R&D. As evidenced by European survey data, about half of European firms that innovate do not conduct R&D (Huang, Arundel, Hollanders, 2007). The share of firms that innovate without doing R&D is likely even higher in developing countries, where much of the early challenge is to deal with existing 'bottlenecks' (Sutz, 2012) or 'weak innovation systems' (UNU-INTECH, 2005). Policies in developing countries should therefore pay sufficient attention toinnovation in terms of organizations, non-technological innovation and the import of embodied technologies not involving own R&D activities.

4. INNOVATION SURVEYS (CIS, FOLLOWING OSLO MANUAL)

Innovation surveys are conducted to collect information on innovation. Innovation surveys, in Europe represented by the Community Innovation Survey (CIS), follow the Oslo Manual(3rd revision) guidelines how to measure innovation. Innovation surveys ask the performers of innovation (i.e. firms) whether they conduct certain activities that lead to innovation. The definition of innovation, collection methods, survey questions, and data compilations have evolved over the years to improve the quality of statistics and it is closely linked with the evolutionary change in the Oslo Manual (from original to revision 4)[1]. The survey developed for European countries, the CIS, and the Oslo Manual are applied in most of the emerging countries by adapting the questionnaire to the local context while keeping comparability. The degree of modification of the CIS questionnaire essentially depends on the choices of these countries on what they want to find out regarding innovation and innovation policy.

Innovation survey data basically complement existing data on patents, bibliometric indicators and R&D surveys. Hence, the survey basically provides the following information (Mohnen and Mairesse, 2010: 6):

- Indicators of innovation output (such as the introduction of new products and processes, organizational changes and marketing innovations, the percentages of sales due to products new to the firm or new to the market, and the share of products at various stages of the product life-cycle);
- A wider range of innovation expenditures or activities than R&D expenditures (such as the acquisition of patents and licenses, product design, personnel training, trail production, and market analysis);
- Information about the way innovation precedes, such as sources of knowledge, the reasons to innovate and perceived obstacles to innovation.

The Oslo Manual follows the subject approach of survey which is collecting information from the firm level instead of object approach, which collects information on innovation (SPRU study), the number of innovation 'output'. The subject approach collects comprehensive data at the decision making level of the firm allowing to conduct much richer analysis that can be linked to the sectoral statistics and national accounts while the drawback of the subject approach is that it does not distinguish between successful and unsuccessful innovations.

The data obtained from innovation surveys are qualitative, subjective and censored (Mohnen and Mairesse, 2010). The number of variables are censored and selected as samples (unless

[1] Detail history of evolutionary development please see following: for Oslo manual (Gault, 2013) and for Community Innovation survey (Arundel and Smith, 2013).

otherwise it is census) and hence subject to some biases (for example, sector). The information obtained is subjective and the quality of variables may contain errors.

Developing countries' problem of adapting the survey:
Innovation indicators started to being adapted in many developing countries since the 1990s. In fact, implementation of innovation surveys in Latin America is not so different from that of Europe (RYCT) and similar interests were also expressed in Africa as can be seen from a NEPAD study (UNU-INTECH). While the recognition on the importance of innovation was present from the early days, earlier experiences of applying Oslo manual based innovation surveys suffered difficulties in not quite capturing the particularities of developing countries. The Bogota manual, as the result, was produced by Colciencia in response to meet the different ideosyncracy of the Latin American innovation context which were later incorporated in the annex of the third revision of the Oslo manual. Many developing countries are currently trying to start conducting innovation surveys. Most of these countries follow the Oslo manual by adapting the CIS survey to understand the innovation process in the country (Gault, 2013, Crespi and Periano, 2007). Many developing countries question the usefulness of conducting an innovation survey. The reasons are as follows.

High cost and barrier
In developing countries, collecting data is much more difficult due to not having fully equipped and capable statistical offices, who may need to prioritize different demands coming from the government (be it demographic data, household survey data etc). In other words, the opportunity cost of conducting an innovation survey is high, especially compared to developed countries. Some of these countries may need to start from building business registries to have acceptable level of selectivity.

Fitting to its economic and industrial structure?
Furthermore, as innovation surveys were originally designed for the developed countries, survey results may not reflect the actual economic/industrial reality in developing countries. For example, many developing countries have a large informal sector (de Beer et al, 2013, Iizuka et al., 2015, Konte and Ndong, 2012). This means that even with well-developed business registries, the survey can only illustrate a relatively small part of economic activities. Moreover, even if the survey is conducted following the Oslo manual, with guidelines based on the experiences of developed countries, copy-pasting the survey questions would not lead to the output that may serve the needs of policy makers in improving innovation policy. For instance, the industrial structure of many African countries demonstrate the important role played by agriculture in its contribution to economic activities as well as in creating employment (see table in Iizuka et al, 2015). The CIS and Oslo manual currently cover the manufacturing, service and mining and quarry sectors, however, they do not cover agriculture. Hence, survey methodologies that can capture the innovation process in agriculture is needed. In fact, nascent attempts are made in Agriculture by ANNI in Uruguay where they have surveyed the agricultural sector (Aboal et al, 2015). While these attempts were already being made, it could take a rather long time to standardize survey questions to be shared among countries with a large agricultural sector.

Finer adjustment to how developing countries innovate
In the similar vein, some of the questions typically used for innovation surveys may require an adaptation to the reality of developing countries. For instance, the minimum size of the firm to be surveyed would be much smaller in developing countries. The definition and type

of 'innovative activities' in developing countries should include (Sutz, 2012) for instance, acquisitions of embodied technology (equipment), minor or incremental changes made in production process, organizational changes, and intentions to conduct innovation.

Sutz (2012) also states that as many developing countries are still yet to develop Innovation capabilities, the investments made in building the system should also be considered as part of expenditure. This would involve the investments in human resources, linkages, quality assurance systems and use of ICTs (Intarakumnerd, 2007).

How to make sense of the survey results to relevant actors?

Resources are often limited in developing countries but some countries manage to conduct an innovation survey. While this is good news, countries are often confronted with other problems: applying the information for policy purposes. For example, the existing survey conducted by the ANII (Uruguay), demonstrated a very low share of policy makers had actually used survey results for innovation policies (Baptista et al. 2009). A comparative study among Chile, Colombia and Uruguay showed similar tendencies. Possible reasons for low policy use of innovation survey data are as follows. First, innovation survey data only become available after some time (results becomes available one year (if not more) later than the reference year of the survey) so it is likely that these data are not perceived as 'up to date' enough to be readily used for policy making. Second, these data may not be elaborated in the way policy makers can comprehend and use them correctly. Third, restricted availability and accessibility of data (in particular micro or firm level data) may cause an insufficient analysis of the data (this may lead to the question of making data publically accessible taking into account confidentiality issues).

Furthermore, considering the globalization of activities through extending value chains, many developing countries are technologically catching up through entering markets by producing goods at lower prices. In many developing countries less patentable 'process' incremental and organizational innovations would be more prevalent than radical innovations through active investments in R&D.

Thirdly, the structural composition of developing countries should be considered carefully. The trends of developing countries are diverse. For instance, many African and Latin American countries have industrial structures with less diversity and reliance on natural resources while some had experienced strong growth in services. In addition to above differences, the size of the informal economy is also substantial in these countries (de Beer et al, 2013, among others).

For instance, in developing countries where most of the countries do not conduct R&D to innovate (Gault, 2010), "learning" and "problem solving" are important parts of the innovation process. In developing countries, due to the under provision of various basic infrastructures (physical, legal, institutional), much of firms' innovative efforts are being made in overcoming existing 'problems' which are not directly considered as 'innovation' in a developed country context (Sutz, 2012). Hence, more firms conduct innovations that do not have R&D nor involve new technology in developing countries. The above example demonstrates the presence of the gap in what constitutes 'problem solving' and 'learning' in different context even though the same word is used.

5. CASE OF COMPOSITE INDICATOR

Due to an increasing availability and accessibility of diverse sets of data, composite indicators are more and more easily constructed and used in the policy domain. Composite indicators summarize individual indicators by compiling these into a single index. Several composite indicators to measure 'innovation' capacity at country level emerged recently such as the Global Innovation Index (WIPO), Global Competitiveness Report (World Economic Forum) and the European Innovation Scoreboard (European Commission) just to name a few. The use of composite indicators became prevalent in the 2000s. Due to the ability of a composite indicator to summarize multidimensional characteristics of complex ideas such as innovation and its facility to communicate and compare results, composite indicators are a powerful policy tool by creating a policy narrative (Saltelli, 2007) while caution of these users are well expressed by numerous experts (OECD JRC handbook, 2008, Freudenberg, 2003, Nardo and Saisana, 200x, Foray and Hollanders, 2015, Edquist and Zabala-Iturriagagoitia 2015, Shibany and Streicher, 2008; Adam, 2014 amongst others).

While the intended use of composite indicators is to grasp overview and monitor progress for policy purposes (Grupp and Mogee, 2004), the ranking table of indices is easily being politicized and a powerful tool (for policy makers to dialogue with the public/budget officers) to mobilize the policy agenda by creating a narrative (Saltelli, 2007). While policy makers can use composite indicators to comply with 'evidence based' and 'participatory' policy making requirements that are increasingly being presented, many users of these indicators may not have a clear understanding how these indicators are constructed and the limitations in what can be interpreted. This potentially creates the information and knowledge asymmetry between different types of users (e.g. policy makers, academics, journalist, lay citizen) making both intentional and unintentional misuses of composite indicators possible. Composite indicators, by definition, give a relative performance benchmark between countries. A common mistake is that a decline in rank performance is interpreted as a real performance decline whereas in most cases a lower rank is not the result of a declining performance but of other countries' performance improving at an even faster rate.

This is especially true in identifying policy prescriptions using composite indicators. For instance, the composite indicator should be analyzed with alternative data sets to understand in detail about the country in disaggregate form. E.g., two countries can have identical scores for their composite indicator hiding significant differences on some of the underlying pillars with one country clearly performing better on inputs in the innovation process like human resources and the other country on outputs of the innovation process like exporting knowledge-intensive products or selling technological knowledge (technological balance of payments). For such diagnostic purposes, innovation survey data and other available information on R&D, human resources and economic indicators become useful. In fact, several studies (OECD/JRC, 2008, Nardo and Saisana, 2008, Adams, 2014) clearly stated, composite indicators are good in evaluating a country's innovation performance in relative terms compared to other countries on selected indicators for well-defined purposes; however, these indicators are not well fitted to conduct policy analysis for evaluating and monitoring the implemented innovation policies in detail.

6. CONCLUSION: TOWARDS DISCUSSION

In interpreting innovation indicators and composite indicators, one needs to take into account that:

-Indicators are a qualitative construct, not a scientific measurement; hence its interpretation requires utmost care in understanding its underlying theoretical/conceptual constructs and selection of data;

-Useful measurements are unique to each county; hence knowing the industrial structure of the country can clarify what are the information needed;

-More is not always better, all the elements need to be studied in the context and in proportion, coordination with other sector/activities and in sequence (order); (good interpretation requires to understand the context in which the indicator is used (be it a country, industrial structure or sector);

-No one prescription fits all, identify clear purpose of use; indicators are products of difficult compromises and one needs to know what has been compromised;

-Indicators are not written on stone, it will change with the changing reality; hence constant discussion, amendments and updates are expected. This is clear from series of revisions that has taken place in e.g. the Frascati and Oslo manuals.

REFERENCES

Aboal, D., Mondelli, M.P., Vairo, M., (2015) Innovation and productivity in Agricultural Firms: Evidence from a Farm-level innovation Survey, Mimeo.

Adam, F.(2014) Measuring National Innovation Performance : The Innovation Union Scoreboard Revisited, Springer Briefs in Economics, Hidelberg, Springer.

Baptista, B., Bernheim, R.,Garce, A, Hernandez, E., (2009) Consulta a Tomadores de Decision en Politicas Publicas de Ciencia, Tecnologia e Innovacion sobre sus Fuentes de Informacion: Informe Regional, Montevideo, Uruguay, ANII.

Crespi, G. and Peirano, F. (2007)Measuring Innovation in Latin America: what we did, where we are and what we want to do, Paper prepared for the Conference on Micro Evidence on Innovation in Developing countries, Maastricht, the Netherlands, UNU-MERIT.

De Beer, J., Fu, K., Wunsch-Vincent, S., (2013) The informal economy, innovation and intellectual property—Concepts, metrics and policy considerations. Economics Research Working Paper No. 10, Geneva, Switzerland, WIPO.

Edquist, C and Zasbala-Iturriagagoita, Jon Mikel (2015) The innovation Union Scoreborad is flawed: The case of Sweden-not the innovation leader of the EU-updated version, Papers in Innovation studies paper no. 2015/27, CIRCLE, Lund University.

Foray, D and Hollanders, H (2015) An assessment of the Innovation Union Scoreboard as a tool to analyze national innovation capacities: The case of Switzerland, Research Evaluation, online version: doi: 10.1093/reseval/rvu036: pp. 1-16.

Freeman, C., Soete, L., (2009) Developing science, technology and innovation indicators: What we can learn from the past, Research Policy, 38: 538-589.

Freudenberg, M., (2003) Composite Indicators of Country Performance: a critical assessment, OECD Science, Technology and Industry Working papers, 2003/16, OECD.

Gault, F. (2010) Innovation Strategies for a Global Economy: Development, Implementation, measurement and Management, Cheltenham, Edward Elgar.

Gault, F (ed.) (2013) Handbook of Innovation Indicators and Measurement, Cheltenham, Edward Elgar.

Grupp, H. and Mogee, M.E. (2004) Indicators for national science and technology policy: how robust are composite indicators?, Research Policy, 33: 1373-1384.

Hatzichronoglou, T. (1997), "Revision of the High-Technology Sector and Product Classification", OECD Science, Technology and Industry Working Papers, No. 1997/02. doi: 10.1787/134337307632.

Hirsch Kreinsen, H., Schwinge, I., (2014)Knowledge-Intensive Entrepreneurship in Low-Tech Industries, Cheltenham, Edward Elgaer.

Huang, Can, Anthony Arundel & Hugo Hollanders, 2007, Non-R&D Innovation of Manufacturing Firms: Theory and Evidence from the Third European Community Innovation Survey, Innovation Watch. Brussels: European Commission.

Iizuka, Michiko, Philippe Mawoko, & Fred Gault (eds.) 2015, Innovation for Development in Southern and Eastern Africa: Challenges for Promoting ST&I Policy UNU-MERIT POLICY BRIEF, 8pages.

Intarakumnerd, Patarapong,(2007) Measuring innovation in Catching up Economies: An experience from Thailand. Paper obtained from internet accessed February, 2016.

Konte, A. and Ndong, M.(2012) The Informal ICT sector and innovation Processes in Senegal, UNU Working paper 2012-009, Maastricht, UNU-MERIT.

Mendonca, S. and N. von Tunzelmann (2004). Brave Old World: Accounting for "high tech" knowledge in "low tech" industries. DRUID Summer Conference 2004, Elsinore, Denmark, June 14-16.

Mairesse, J., Mohnen, P., (2010) Using Innovations Surveys for Econometric Analysis, NBER Working Paper Series 15857, Cambridge, USA, NBER.

Nardo, M. Saisana, M., (2008) OECD/JRC Handbook on constructing composite indicators. Putting theory into practice.

OECD/ JRC (2008) Handbook on Constructing Composite Indicators: methodology and user guide, Paris, OECD

OECD (2012) Recommendation of the Council on Regulatory Policy and Governance, Paris,OECD.

OECD (2011) ISIC REV. 3 Technology Intensity Definition: Classification of manufacturing industries into categories based on R&D intensities, OECD Directorate for Science, Technology and Industry Economic Analysis and Statistics Division, Paris, OECD.

von Tunzelmann, N. and V. Acha (2005), Innovation in 'low-tech' industries. Jan Fagerberg, David Mowery, Richard Nelson (eds), The Oxford Handbook of Innovation, New York: Oxford University Press: 407-432.

Saltelli, A., (2007) Composite Indicators Between Analysis and Advocacy, Social Indicators Research, 81: 65-77.

Saisana, M. Saltelli, A. and Tarantola, S. (2005) Uncertainty and sensitivity analysis techniques as tools for the quality assessment of composite indicators, J.R. Statist. Soc.A 168 (2): 307-323.

Schibany, A. and Streicher, G. (2008) How not to compare innovation performance: A critical assessment of European Innovation Scoreboard, Institute of Technology and Regional Policy-InTeREg

Smith, K., (2005) Measuring Innovation, In Fagerberg, J., Mowery, D., Nelson R.R., Edited, The Oxford Handbook of Innovation, Oxford: Oxford University Press.

Sutz, J (2012) Measuring innovation in developing countries: some suggestions to achieve more accurate and useful indicators, International Journal of Technological Learning Innovation and Development, 5 (1/2): 40-57.

UNU-INTECH (2005) Measuring Innovation: Making innovation surveys work for developing countries, Technology Policy Briefs, 4 (1) 11 pages.

Analysing innovation policy indicators through a functional approach: the aeronautic industry case1

Carolina Resende Haddad*, Mauricio Uriona Maldonado**

*carol.resende.haddad@gmail.com
Department of Industrial and Systems Engineering, Federal University of Santa Catarina, Campus Universitario - Trindade, Florianopolis, 88040-270 (Brazil)

** m.uriona@ufsc.br
Department of Industrial and Systems Engineering, Federal University of Santa Catarina, Campus Universitario - Trindade, Florianopolis, 88040-270 (Brazil)

ABSTRACT
Developing countries face different problems than developed countries and the use of the same indicator to evaluate and compare both regions can lead to misleading conclusions. Traditional indicators, such as R&D and patents may not capture the whole dynamic of a system, as they are used to compare systems focusing on its current structure. Many authors have been discussing the processes underlying industry transformation, innovation, and economic growth to access a system performance, i.e. the functions of innovation systems. Therefore, the purpose of this paper is to analyze these functions as indicators to measure the performance of the system in order to identify policy issues. In order to do that, we analyze the case of the aeronautic sectoral system of innovation of a region in Brazil. The functional approach helped us to better capture the dynamic of the system, by not restricting our analysis to the system's structure.

INTRODUCTION
In order to develop public policies to stimulate innovation towards local, regional or national needs, we need to understand how innovative a system is (Grupp & Schubert, 2010), which presupposes the ability to measure innovation. Therefore, many policymakers have discussed the development of indicators to better capture innovation activities (Gaut, 2013; Lee, 2015). Most parts of the typical innovation measurement tools are based on the linear model of innovation, i.e., on one or two indicators, such as patents and R&D (Mahroum & Al-Saleh, 2013). However, as argued by Archibugi, Denni and Filippetti (2009), many innovation indicators are not helpful for measuring innovation, as they do not reflect the innovation factors that distinguish different countries.

Lepori, Barré and Filliatreau (2008) pointed out that in the past decades there was the increase and diversification of STI indicators and innovation measurement in terms of analysis, types, consumers, and users, mainly due to the increasing complexity of the systems. For example, we can mention the use of composite innovation indicators (Grupp & Schubert, 2010). More recently, Mahroum and Al-Saleh (2013) proposed a measurement tool called the "Innovation Efficacy Index", which considers five functions of the "innovation through adoption" process

[1] This work was supported by the Industry Federation of Santa Catarina (FIESC).

(accessing, anchoring, diffusing, creating and exploiting innovations) in order to understand cross-countries differences in innovation performance.

In this functionality line, many authors have been discussing the processes underlying industry transformation, innovation, and economic growth in order to evaluate the dynamic of the innovation system (Jacobsson, & Johnson, 2000; Liu, & White, 2001; Johnson, 2001; Hekkert *et al*., 2007; Bergek *et al*., 2008). These processes were labelled functions of innovation systems (Bergek *et al*., 2008), which are the activities that take place in this system in order to generate technological change and disseminate innovations (Hekkert *et al*., 2007). In this sense, the purpose of this paper is to analyze the functions of the innovation system as indicators to measure the performance of an emerging system of innovation of a developing country in order to identify policy issues.

We analyse the case of the aeronautic sectoral system of innovation in Santa Catarina State, Brazil. As pointed out by Hekkert *et al*. (2007), the functional approach (i) allows the comparison between innovation systems with different backgrounds; (ii) allows a systematic method of mapping the determinants of innovation; and (iii) allows the formulation of a set of policies that should be the target of the innovation system and the tools to achieve this target.

FUNCTIONS OF THE INNOVATION SYSTEM

The functional approach are related to the character and the interaction between the components of the innovation system (agents, networks and institutions) (Hekkert, & Negro, 2009). It was originally developed for Technology Innovation Systems, focused mainly on renewable energies (Jacobsson & Bergek, 2011; Negro, Hekkert & Smits, 2007). Other works extrapolated the renewable energy TIS and analyzed other IS, e.g. the ceramic tile Sectoral Innovation System (Gabaldón-Estevan; Hekkert, 2013).

Within the many attempts to identify functions, we will use the functions proposed by Hekkert *et al*. (2007), which are:

- Entrepreneurial activities: new entrants that identify an opportunity in the market and companies that diversify their business strategies;
- Knowledge development: mechanisms of learning, encompassing “learning by searching” and “learning by doing”;
- Knowledge diffusion through networks: is the exchange of information between actors in the innovation system;
- Guidance of the search: choose the focus of investments in technology among the options;
- Market formation: is the creation of protected spaces for new technologies, such as the formation of niche markets or by creating favourable tax regimes;
- Resources mobilization: is the allocation of resources, both financial and human capital, for specific technologies;
- Creation of legitimacy/counteracts resistance to change.

Table 1 shows the typical indicators to measure each of the seven functions (Hekkert *et al*., 2007).

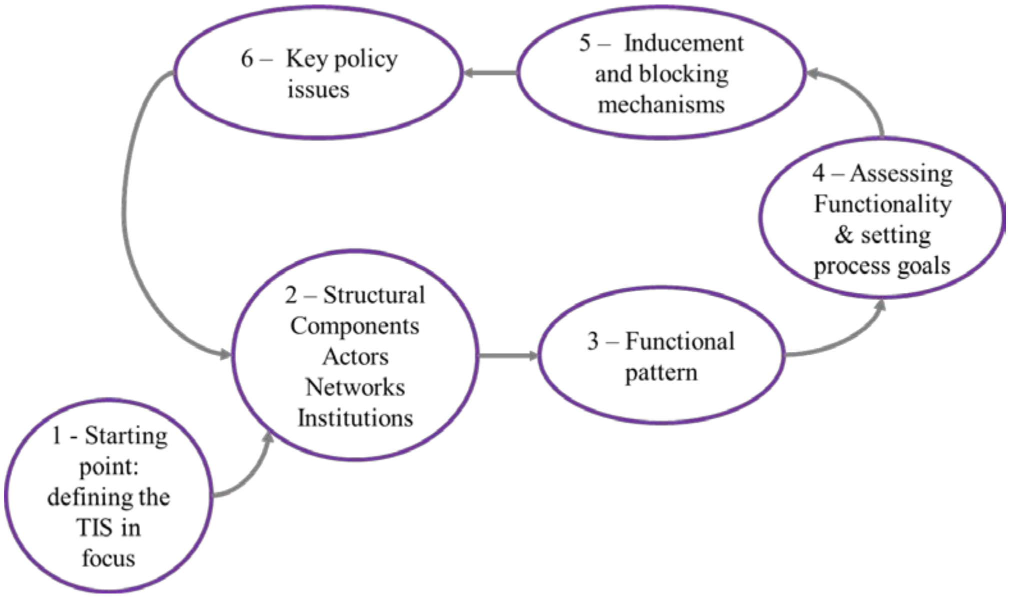
Table 1: typical indicators to measure the Functions of the Innovation System.

Function	**Typical indicators**
Entrepreneurial activities	Number of new entrants; Number of diversification activities; Number of new experiments with a new technology.
Knowledge development	R&D projects over time; Patents; Investments in R&D.
Knowledge diffusion through networks	Number of workshops and conferences on a particular technology The network size and intensity over time.
Guidance of the search	Specific targets set by governments or industries regarding the use of a specific technology; Number of articles in professional journals that raise expectations about new technological development.
Market formation	Number of niche markets that have been introduced; Specific tax regimes for new technologies; New environmental standards that improve the chances for new environmental technologies.
Resources mobilization	Funds made available for long-term R&D programs set up by industry or government to develop specific technological knowledge; Funds made available to allow testing of new technologies in niche experiments; Perception of the actors regarding the access to sufficient resources.
Creation of legitimacy/counteracts resistance to change	Rise and growth of interest groups; Lobby actions.

Source: adapted from Hekkert *et al*. (2007).

In order to analyse the SI and help policy makers in the selection and prioritization of public policies, Bergek *et al*. (2008) proposed an analytical scheme that allows accessing the performance of the system and identifying the aspects that are influencing this performance:

Figure 1: Scheme of analysis.



Source: adapted from Bergek *et al*. (2008).

METHOD

Based on the main aim of our paper, we followed the six steps proposed by Bergek *et al*. (2008) to analyse the functional dynamics of the aeronautic industry in Santa Catarina's State:

- Step 1: we defined the focus of the Sectoral System of Innovation;
- Step 2: we identified the structural components of the innovation system;
- Step 3: we mapped the functional pattern of the system considering the seven functions of the innovation system proposed by Hekkert *et al*. (2007). In this step, we collected the data for document analysis and expert's interviews;
- Step 4: we accessed the functionality of the system by analysing its phase of development and set the final process goal;
- Step 5: we identified the system's inducing and blocking mechanism;
- Step 6: we specified the key policy issues concerning the aeronautic industry final process goal.

We conducted expert panels in order to collect data to steps 4 to 6.

THE AERONAUTIC INDUSTRY CASE

The system that will be the focus of this paper is the aeronautic industry of Santa Catarina's State (SC), Brazil. A survey made by the Industry Federation of Santa Catarina (FIESC, 2013) identified that this Sectoral Innovation System was diagnosed to be in an emerging level.

Step 1 – Define the system's focus

To set the focus of the Innovation Sector Aeronautic system in SC, we used the Brazilian National Classification of Economic Activities (CNAE) number 30, subsections number 30.4, 30.5 e 30.9, which are specifically related to the aeronautic industry.

Step 2 – Define the structural components

Due to the emerging level of the industry analysis, it was difficult to clearly identify those actors, networks and institutions that strongly influence the aeronautic industry. Thus, we will discuss in general terms those who are nowadays are present at this stage of the industry.

In term of actors, SC is characterized by the presence of few companies in the sector. FIESC is an active actor in the industry, as it represent all the industries in the state. We can also mention as actors, labour unions and regulatory agencies. Some universities and regional research institutes act as actors in the system, but they have no prominent role in the aeronautic system yet.

Considering the aeronautic system networks, we identified the relationship between suppliers of different levels of the supply chain and between companies and labour unions. The university-enterprise network little influences the system nowadays.

We observed a lack of institutions of interest in the system, especially because it is still in an emerging stage. The National Civil Aviation Agency (ANAC) is one of the agencies that regulate the sector in Brazil.

Step 3 – Analyse the functionality

We describe the main findings of the aeronautic system functionality below.

Entrepreneurial activities

Number of new entrants
- It is expected that Novaer, a new entrant that produces small aircraft, will install a factory in Lages, in the South of SC;
- Lack of entrepreneurs and companies specialized in the production of high value-added products for the aeronautic industry.

Number of diversification activities
- There is little diversification activities in the aeronautic industry.

Number of new experiments with a new technology
- New experiments are being developed by entrepreneurs in small aircraft;
- Lack of experiments in fuselage.

Knowledge development

R&D projects over time
- Knowledge is being developed through "learning by doing" in the construction of small aircraft;

- Researches are being developed in kits for light aircraft;

Patents

- A search in the Brazilian National Institute of Industrial Property (INPI), we could find 49 patents registered with "aeronautic" in the title (INPI, 2016).

Investments in R&D

- Novaer Craft (2015) plans to install an engineering center in Florianopolis in order to develop research in the aeronautical sector.

Knowledge diffusion through networks

Number of workshops and conferences on a particular technology

- The Development Committee of the Aeronautical Industry is responsible for the diffusion of knowledge in the sector by promoting courses focused on the industry (FIESC, 2016);
- Some actors in the IS are participating in international conferences and workshops, such as the SUN'n FUN in the United States, which help to approximate regional companies to international ones.

The network size and intensity over time

- Lack of interaction between the university, government and industries, which could be better explored to diffuse knowledge through the system networks.
- The *Brazilian Aerospace Cluster* Project, developed by the Brazilian Spatial Agency (AEB, 2014), aims at creating an aeronautic industry cluster in SC.

Guidance of the search

Specific targets set by governments or industries regarding the use of a specific technology

- The Development Program for Certification to Small-Sized Aircraft (iBR2020), prepared by ANAC (2015a), develops projects focused on small aircraft in order to make then more prepared to succeed when subjected to certification;
- Another factor that influences the direction of search are the standards for the aerospace industry based on ISO 9001, from the AS/EN 9100 series;
- Other certifications related to the sector are ISO 14.001 (environmental management), Management System of Occupational Health and Safety - OHSAS 18.001 (2015), the Type-certificate for aircraft, engines and propellers, and the Aircraft Certificate in transport category (ANAC, 2015b);
- The *Federal Aviation Administration* (FAA), from the EUA, and the *European Aviation Safety Agency* (EASA) regulate the civil aviation and influence the guidance of the search;
- The Development Committee of the Aeronautical Industry influences the direction of search since it promotes discussions about the industry's technological guidelines in the state.

Number of articles in professional journals that raise expectations about new technological development

- We did not find any aspect related to this indicator.

Para seguir leyendo haga click aquí