

XIII Encontro Nacional de Pesquisa em Ciência da Informação - XIII ENANCIB 2012
GT 11: Informação e Saúde

A IMPORTÂNCIA DA PADRONIZAÇÃO NA INFORMETRIA: UM ESTUDO
EXPLORATÓRIO NA ÁREA DE SAÚDE PÚBLICA

Comunicação Oral

Rosany Bochner - Fundação Oswaldo Cruz – Fiocruz
Maria Simone de Menezes Alencar - Fundação Oswaldo Cruz – Fiocruz
Viviane Santos de Oliveira Veiga - Fundação Oswaldo Cruz – Fiocruz
Rejane Ramos Machado - Fundação Oswaldo Cruz – Fiocruz
rosany@icict.fiocruz.br

Resumo

Os estudos informétricos referem-se aos aspectos quantitativos da informação em qualquer formato, incluindo as bases de dados. A robustez desses estudos depende da qualidade dos dados, que podem afetar não somente os estudos de colaboração, evolução de redes de cooperação e a delimitação de campos científicos, mas também a visibilidade de instituições e o ranking de organizações de pesquisa. Este estudo objetiva evidenciar o impacto que a falta de padronização causa nos estudos métricos da ciência. Foram recuperados os registros sobre a temática de saúde pública, ambiental e ocupacional publicada por primeiro autor vinculado à instituição localizada no Brasil, durante o período de 2001 a 2011, na *Base Web of Science* via Portal de Periódicos CAPES e importados para o software de mineração de dados VantagePoint. Em análise do campo de nome da instituição dos autores foi possível observar diversas formas de grafia que geram problemas na identificação inequívoca: uso de siglas, abreviaturas, omissão, transposição, erros de digitação, espaçamento, pontuação, letras maiúsculas e minúsculas, tradução para outros idiomas, ou a combinação das anteriores. Neste universo a Fundação Oswaldo Cruz foi a que apresentou o maior número de variações (45), e um incremento de 95% no número de artigos após a unificação da entrada do nome da instituição, passando de 353 para 689 artigos, seguida da Universidade Estadual Paulista - UNESP (21) e da Comissão Nacional de Energia Nuclear - CNEN (16). A filiação impacta a análise da produção científica por instituição e pode influenciar também no ranking de produção de cada autor em particular. Para evitar equívocos deve-se estabelecer procedimentos padronizados de vinculação profissional no momento em que o autor está inserindo seus dados no artigo, além de conscientização e consenso por parte de todos os atores do processo de comunicação científica.

Abstract

Informetric studies are related to the quantitative information in any form, including databases. The robustness of these studies depends on the quality of data, which can affect not only the collaborative studies, development of cooperation networks and the delimitation of scientific fields, but also the visibility of institutions and ranking of research organizations. This study aims at identifying the impact that the lack of standardization concerned in metrical studies of science. We retrieved the records on the subject of Public, Environmental & Occupational Health published by the first author from the institution located in Brazil, during the period 2001 to 2011, at the Base Web of Science via CAPES Periodicals Portal and imported for data mining software VantagePoint. In analysis of the field of the author's institution was possible to observe various forms of spelling that create problems in identifying unequivocal: the use of acronyms, abbreviations, omission, transposition, typos, spacing, punctuation, capitalization, translation into other languages, or a combination of the above. In this universe the Fundação Oswaldo Cruz was the one with the largest number of variations (45), and an increase of 95% in the number of articles after the unification of entering the name of the institution, from 353 to 689 items, then the Universidade Estadual Paulista - UNESP (21) and the Comissão Nacional de Energia Nuclear - CNEN (16). The affiliation impacts the analysis of scientific production by the institution and may also influence the ranking of production of each particular author. To avoid misunderstandings should be established standard procedures to linking professional at the time the author is inserting your data in the article, as well as awareness and consensus by all the actors in the scholarly communication process.

1. INTRODUÇÃO

O processo de comunicação científica definido por Garvey (1979) consiste em uma sucessão de etapas que vão desde o momento em que se concebe uma idéia para pesquisa até que essa se torne pública através da publicação científica. Uma vez que o conhecimento científico só pode se considerado existente se além de publicado puder ser recuperado em diversos contextos (ROTHER, 2007), o processo inicialmente definido por Garvey poderia ser complementado com a indexação e recuperação em bases de dados.

Nesse contexto, a padronização dos dados faz-se presente. A revisão da literatura aponta para essa questão desde o início dos anos 80, quando Williams e Lannom (1981) e Piternick (1982) discutiram a falta de padronização dos títulos de periódicos. Meneghini (1995) verificou “pouco caso por parte dos pesquisadores brasileiros” na forma de indicar sua filiação. Rother afirma que:

“A normalização é um dos critérios para registro e inclusão em bases de dados. Dentre os principais aspectos que precisam ser considerados na normalização de uma publicação estão os intrínsecos que se referem à padronização dos resumos, descritores e identidade dos autores” (2007, p. 225).

Glänzel (1996) discute a necessidade de padronização na pesquisa bibliométrica e afirma que o desenvolvimento de padrões é necessário para garantir a validade dos métodos bibliométricos, melhorar a confiabilidade de seus resultados e tornar seus dados compatíveis.

Os estudos informétricos segundo Macias-Chapula (1998) referem-se aos aspectos quantitativos da informação em qualquer formato, incluindo as bases de dados. A robustez desses estudos depende da qualidade dos dados, o que inclui sua completeza, correção, disponibilidade, interpretabilidade e consistência (GALVEZ; MOYA-ANEGÓN, 2007). Essas dimensões devem ser consideradas no início do processo, no momento em que a informação referente a cada campo é gerada, ou seja, quando da submissão do artigo aos periódicos científicos. Nesse sentido, deve-se considerar que essa responsabilidade não é só do autor, cabendo também aos editores no que se refere às normas de submissão. Se por um lado os editores não têm domínio sobre as diferentes formas adotadas pelo autor para grafar seu nome, por outro podem definir critérios para sua filiação.

Quando a qualidade dos dados não é garantida no início do processo, tem-se como resultado bases de dados com erros e inconsistências, que podem afetar não somente os estudos

de colaboração, evolução de redes de cooperação e a delimitação de campos científicos, mas também a visibilidade de instituições e o ranking de organizações de pesquisa (GALVEZ; MOYA-ANEGÓN, 2006).

A produção científica do campo da saúde, que é o objeto desta análise, é um domínio privilegiado para discutir a importância da padronização nos estudos informétricos, especialmente pelo seu volume e representatividade na literatura científica mundial registrada nas principais bases de dados. Não seria surpresa, portanto, que as questões relacionadas à padronização não expressassem relevo e importância particularmente das autorias e vinculações institucionais, seja na variedade de vínculos, nas parcerias instituídas, bem como na questão lingüística. Nessa linha de raciocínio, o objetivo da análise aqui apresentada, de caráter preliminar, é evidenciar o impacto da falta de padronização a partir da recuperação de informação nas bases de dados nos estudos informétricos.

2. METODOLOGIA

Para realizar a análise aqui relatada foi utilizada a base de dados *Web of Science* da Thompson Reuters através do Portal de Periódicos da CAPES. Enquanto fonte de informação, ela se caracteriza por ser multidisciplinar, com cobertura retrospectiva desde 1900 em algumas áreas do conhecimento. Tradicional base de estudos bibliométricos por sua abrangência, a *Web of Science* permite a busca por categorias de assunto (PORTER; CUNNINGHAM, 2005). As categorias de assunto são um esquema classificatório que organizam os periódicos por áreas do conhecimento. Esta estratégia foi utilizada para recuperar os registros sobre a temática de saúde pública, ambiental e ocupacional publicada por pelo menos um autor vinculado à instituição localizada no Brasil, durante o período de 2001 a 2011. O resultado da busca foi refinado de forma a recuperar apenas os artigos de periódicos, eliminando outras tipologias tais como editoriais, *proceedings*, *reviews*, etc. Os artigos de periódicos são, segundo Ziman (1968), a principal evidência da produção de conhecimento em uma área de pesquisa, dada sua certificação científica fornecida tanto pelos editores como pela revisão dos pares.

Os metadados dos artigos foram recuperados em formato texto a partir da *Web of Science* para tratamento em software de mineração de dados. Os registros recuperados foram importados para o software VantagePoint com vistas a tratamento e análise. Esse software é uma solução

para mineração de textos desenvolvida pelo *Georgia Institute of Technology*, comercializada pela empresa americana *Search Technology* e licenciada pelo ICICT/Fiocruz desde 2007. Esta ferramenta permite análise de campos da base de dados individualmente, faz correção entre campos e permite a construção de matrizes intra e inter campos.

Após a importação dos dados, um novo refinamento foi realizado, reduzindo o universo da pesquisa para os artigos cuja filiação do primeiro autor fosse brasileira, de forma a otimizar a identificação inequívoca dos nomes das instituições.

Observou-se a existência de artigos com o mesmo título, havendo a possibilidade de serem duplicatas (mesmo artigo), resultante de erros de indexação na base de dados. Para a checagem desta possibilidade, foi utilizada uma ferramenta semi-automatizada de detecção de duplicadas disponível no software de mineração de dados, na qual são definidos os campos que precisam ser exatos para que se proceda à remoção da duplicata. Foram definidos os seguintes campos: título, autores, periódico e ano de publicação. O passo seguinte foi a comparação entre os registros que não foram eliminados automaticamente. Para tanto foi utilizada a planilha eletrônica Excel, que possui a função “EXATO”, capaz de verificar se duas sequências de caracteres são exatamente iguais, diferenciando letras maiúsculas de minúsculas.

Com o arquivo de registros sem duplicatas, a etapa seguinte foi a padronização do nome das instituições do primeiro autor. Para tal, utilizou-se uma ferramenta de limpeza disponível no software que permite a consolidação de variações de entradas em um dado campo. O processo automatizado usa um algoritmo inteligente que cruza variedades de um mesmo termo, como pontuação, plural, hífen, etc. No entanto, como essa ferramenta foi programada para o idioma inglês, é preciso utilizá-la com muito cuidado e atenção para que não ocorram junções indevidas.

Após o término da padronização, utilizou-se o próprio software VantagePoint para a geração de listas e matrizes para as análises subsequentes, em relação a alteração do posicionamento das instituições com o maior número de publicações. Os resultados e discussões que se seguem foram baseados em dados coletados a partir da estruturação da informação no programa de mineração de dados.

3. RESULTADOS E DISCUSSÃO

A busca na base *Web of Science* inicialmente recuperou 6.608 artigos publicados relacionados à categoria de assunto “*Public, Environmental & Occupational Health*” com algum autor afiliado em instituição localizada no Brasil, no período de 2001 a 2011. Após o refino para o primeiro autor ser de instituição brasileira, o universo da pesquisa foi reduzido a 5.963 artigos.

Neste conjunto foram identificados 16 registros com títulos duplicados. Após a checagem com a ferramenta de remoção de duplicatas do software VantagePoint, ainda resultaram quatro documentos com títulos duplicados. Observou-se que havia diferença de indexação dos autores e do ano. Em consulta na fonte original do artigo constatou-se que em dois casos o ano de publicação estava diferente do original e em outros dois havia problema na ordem de indexação do nome dos autores, bem como erros de grafia. Estes tipos de problemas foram assinalados por Hood e Wilson (2003) ao analisar a falta de consistência de dados eletrônicos. Nesse sentido, citam o trabalho de Pao (1989), no qual este autor caracteriza nove tipos de erros: adição, omissão, transposição, digitação, espaçamento, pontuação, letras maiúsculas e minúsculas, nomes compostos ou a combinação dos anteriores. Esse autor também encontra inconsistência no campo de data. Wolfram et al. (1990 apud HOOD e WILSON, 2003) observam que problemas de data de publicação abrangem cerca de dois anos. Dados dessa pesquisa corroboram essa afirmativa, o que pode ser explicado pela diferença de datas entre o aceite do trabalho, muitas vezes com pré-print online, e a publicação propriamente dita.

A eliminação das duplicadas resultou em um arquivo com 5.947 registros de artigos de periódicos com o primeiro autor vinculado a uma instituição brasileira. Em uma primeira análise do campo onde consta o registro da instituição dos autores foi possível observar diversas formas de grafia que geram problemas na identificação inequívoca: uso de siglas, abreviaturas, omissão, transposição, erros de digitação, espaçamento, pontuação, letras maiúsculas e minúsculas, tradução para outros idiomas, ou a combinação das anteriores. Em suma, alguns dos problemas aqui assinalados para filiação são similares aos citados anteriormente para a grafia de autores.

Diferentes variações de nomes para uma mesma instituição foram observadas. O Quadro 1 apresenta essas variações para as 20 instituições com maior número de artigos inseridos na base *Web of Science* no período pesquisado. É importante observar que o número de artigos atribuídos a cada instituição pode ser alterado de forma substancial com a uniformização do

nome das instituições. Pode-se verificar que a Fundação Oswaldo Cruz - Fiocruz foi a que apresentou o maior número de variações (45), seguida da Universidade Estadual Paulista - UNESP (21) e da Comissão Nacional de Energia Nuclear - CNEN (16).

A Tabela 1 apresenta a distribuição do número de artigos antes e depois da padronização do campo “Filiação do primeiro autor”, classificação e percentual de variação para as 20 instituições brasileiras que mais apresentaram artigos na base *Web of Science* relacionados à categoria de assunto “*Public, Environmental & Occupational Health*”, no período de 2001 a 2011.

No caso da Fiocruz, apesar de manter sua classificação antes e depois da padronização do campo (2º lugar), as diferentes formas de se representar acabou gerando um incremento de 95% no número de artigos após a unificação da entrada do nome da instituição do primeiro autor, passando de 353 para 689 artigos (Tabela 1).

Instituições de âmbito federal como o CNEN e a Fiocruz, ou mesmo estadual como a UNESP, têm maior probabilidade de enfrentar problemas como este, uma vez que possuem várias unidades. Fato este destacado por Van Raan (2005) ao afirmar que em grandes organizações de pesquisa há dificuldade em distinguir seus múltiplos institutos. A UNESP ao passar de 65 para 130 artigos após a padronização subiu uma classificação, de 12º para 11º, refletindo um aumento de 100%. No caso da CNEN, embora na área específica deste estudo, Saúde Pública, tenha pouca representatividade, a falta de padronização refletiu em um aumento de 272% no número de artigos, passando de 18 para 67 ao se unificar suas diversas unidades e diferentes grafias. É interessante observar que a CNEM que estava na 20ª posição, com a padronização passa para a 14ª (Tabela 1).

A filiação não impacta somente a análise da produção científica por instituição, universidade, estado, etc., pode influenciar também na produção de cada autor em particular, uma vez que é utilizada como um critério no momento de separar ou agrupar nomes de autores. Meneghini (1995) menciona que:

“a localização correta e rápida desses dados é importante não só para fins bibliográficos de busca de informação de resultados e dados de pesquisa, mas também para obtenção de indicadores que se tornam cada vez mais importantes em termos de política científico-tecnológica: publicações e impactos por área, instituição, universidade, estado, etc.” (MENECHINI, 1995, p. 396).

O mesmo autor sugere também que se indique a filiação em português, esta apresentada em ordem de hierarquia crescente, os nomes dos departamentos (laboratório, seção), unidade (faculdade, escola, instituto, centro) e instituição (universidade, instituto de pesquisa). Observa também que não se deve, de forma alguma, utilizar abreviações e siglas.

Nesse sentido algumas instituições têm buscado traçar diretrizes no sentido de padronizar a forma de identificação de seus profissionais em comunicações científicas. A Pró-reitoria da Universidade Federal do Rio de Janeiro segue essa recomendação conforme informação coletada em seu site, preocupada com a “contabilidade das publicações”, orienta:

“Com o intuito de padronizarmos a forma de citarmos o nome de nossa Universidade em nossas publicações científicas, indicamos abaixo o padrão que deve ser adotado daqui para frente em qualquer uma de suas publicações:
Departamento/Programa de XXXXXX, Instituto/Faculdade/Escola/Hospital de YYYYYY, Universidade Federal do Rio de Janeiro, Rio de Janeiro, Brazil
O mais importante é citar por extenso o nome de nossa universidade. Se necessário for, pode ser usado Federal University of Rio de Janeiro.”
(UNIVERSIDADE, 2012)

De forma mais enfática, o Instituto Oswaldo Cruz da Fundação Oswaldo Cruz, editou a Portaria 005/2012-IOC em março deste ano, orientando que todas as comunicações oficiais ou científicas devam trazer tanto o nome do instituto como da fundação, seguindo também a hierarquia sugerida por Meneghini (1995).

Observa-se, no entanto, que as iniciativas são isoladas e não há, por parte dos editores científicos uma preocupação no sentido de padronização. A literatura científica sobre bibliometria é farta em artigos que apresentam métodos matemáticos para automatizar a padronização dos dados tentando corrigir o problema quando este já está instaurado, ou seja, no fim do processo (GALVEZ; MOYA-ANEGÓN, 2006, 2007). Nessa etapa, a padronização e a limpeza da base de dados tornam-se tarefas extremamente trabalhosas, que demandam tempo e exigem expertise e conhecimento especial por parte dos usuários, tanto no que se refere ao domínio das ferramentas utilizadas para a realização dessas atividades, bem como para tomar decisões acerca de suprimir, corrigir e unificar dados.

O ideal é trabalhar com a prevenção desses problemas no início do processo, no momento em que o autor está inserindo seus dados no artigo. Para tanto, os editores dos periódicos têm um papel importante, deveriam criar padrões, ao menos para os campos de autor e filiação, a serem compartilhados por toda a comunidade científica.

Como também existem erros gerados na etapa de indexação dos dados na base, seus produtores, hospedeiros e gestores deveriam também controlar a qualidade da informação que entra na base. Nesse sentido, o uso de tesauro faz-se necessário como forma de padronização de dados. Sua construção não deve prescindir da colaboração da comunidade científica.

4. CONCLUSÃO

O uso de bases de dados está envolvido com muitos problemas que precisam ser superados ou pelo menos gerenciados, a fim de se extrair informações úteis para análises informétricas. Muitos desses problemas surgem do fato de que a maioria dessas bases foi criada para fins de recuperação de informação, relegando aos estudos informétricos uma aplicação secundária.

Atualmente, a informação disponível na base de dados *Web of Science* apresenta erros e inconsistências que podem impactar as análises informétricas. Para sanar esse problema, faz-se necessário a realização de limpeza e unificação dos dados, tarefa essa trabalhosa e dispendiosa em termos de tempo e de ferramenta necessária, e que não está livre de erros, uma vez que o usuário toma decisões de suprimir, corrigir e unificar dados de acordo com seu conhecimento e rigor.

Dessa forma, pode-se afirmar que a produção científica das instituições brasileiras no campo da saúde pública, e a de cada autor em particular, só serão recuperadas das bases de dados e avaliadas devidamente se forem estabelecidos procedimentos padronizados de vinculação profissional no início do processo, ou seja, no momento em que o autor está inserindo seus dados no artigo.

Contudo, cabe salientar que a padronização só terá resultados satisfatórios se houver conscientização e consenso por parte de todos os atores do processo de comunicação científica: o autor, o avaliador, revisor e editor dos periódicos e o produtor, hospedeiro e gestor das bases de dados.

5. REFERÊNCIAS

FUNDAÇÃO Oswaldo Cruz. Instituto Oswaldo Cruz. **Portaria 005/2012-IOC**. 01 de março de 2012.

GALVEZ, Carmen; MOYAN-ANEGÓN, Félix. Standardizing formats of corporate source data. **Scientometrics**, v. 70, n. 1, p. 3-26, 2007.

GALVEZ, Carmen; MOYAN-ANEGÓN, Félix. The unification of institutional addresses applying parametrized finite-state graphs (P-FSG). **Scientometrics**, v. 69, n. 2, p. 323-345, 2006.

GARVEY, W. D. **Communication: the essence of science facilitating information exchange among librarians, scientists, engineers and students**. Oxford: Pergamon Press, 1979. 332p.

GLÄNZEL, W. The need for standards in bibliometric research and technology. **Scientometrics**, v. 35, n. 2, pp 167-176.

HOOD, William W.; WILSON, Concepción S. Informetric studies using databases: opportunities and challenges. **Scientometrics**, v. 58, n. 3, p. 587-608, 2003.

MACIAS-CHAPULA, Cesar A. O papel da informetria e da cienciometria e sua perspectiva nacional e internacional. **Ciência da Informação**, v. 27, n. 2, p. 134-140, 1998.

MENEZHINI, Rogério. Padronização de filiação acadêmico-científica, ou como não ter informações sobre suas publicações perdidas nos bancos de dados nacionais e internacionais. **Química Nova**, v. 18, n. 4, p. 396-397, 1995.

PITERNICK, Anne B. Standardization of journal titles in databases. **Journal of the American Society for Information Science**, v. 33, n. 2, p. 105, 1982.

PORTER, Alan. L.; CUNNINGHAM, Scott. W. **Tech mining: exploiting new technologies for competitive advantage**. Estados Unidos, Hoboken, New Jersey: John Wiley & Sons, 2005, 384 p.

ROTHER, Edna Terezinha. O papel da normalização nas publicações científicas [Editorial]. **Revista Brasileira de Oftalmologia**, v. 66, n. 4, p. 225-226, 2007.

UNIVERSIDADE Federal do Rio de Janeiro. Pró-reitoria de Pós-graduação e Pesquisa. **Publicação Científica**. Disponível em: <http://www.pr2.ufrj.br/site/index.php/pr2/publicacaoCientifica>. Acesso em 11 jul 2012.

VAN RAAN, Anthony F. J. Fatal Attraction: conceptual and methodological problems in the ranking of universities of bibliometric methods. **Scientometrics**, v. 62, n. 1, p. 133-143, 2005.

WEB of Science. Disponível em: <http://apps.webofknowledge.com.ez68.periodicos.capes.gov.br>. Acesso em: 08 jun 2012.

WILLIAMS, Martha E.; LANNOM, Lauranece. Lack of standardization of the journal title data element in databases. **Journal of the American Society for Information Science**, v. 32, n. 1, p. 229-233, 1971.

ZIMAN, John. **Public knowledge**. London: Cambridge University Press, 1968.

Quadro 1: Filiação do autor na forma padronizada e nas diferentes formas encontradas na base de dados, com o respectivo número de artigos.

Filiação do autor padronizada [n° artigos]	Filiação do autor nas formas encontradas na base [n° artigos]
Univ São Paulo [825]	Univ Sao Paulo [744]; USP [64]; FMUSP [7]; FSP USP [2]; Hosp Univ Sao Paulo [2]; FFCLRP USP [1]; HC FMUSP [1]; NUPENS USP [1]; Ribeirao Preto USP [1]; USP EEUSP [1]; USP SP [1]
Fiocruz [689]	Fundacao Oswaldo Cruz [353]; Fiocruz MS [171]; Fdn Oswaldo Cruz [23]; Inst Oswaldo Cruz [22]; Escola Nacl Saude Publ Sergio Arouca [19]; Escola Nacl Saude Publica [17]; Fundacao Oswaldo Cruz Fiocruz [13]; Oswaldo Cruz Fdn [9]; FIOCRUZ [6]; Inst Fernandes Figueira Fiocruz [5]; Fdn Oswaldo Cruz FIOCRUZ [4]; Inst Oswaldo Cruz Fiocruz [4]; Ctr Pesquisas Aggeu Magalhaes [3]; ENSP Fiocruz [3]; CPqAM [2]; Ctr Pesquisas Aggeu Magalhaes FIOCRUZ [2]; FIOCRUZ RJ [2]; Fundacao Oswaldo Cruz [2]; Goncalo Moniz Res Ctr [2]; Oswaldo Cruz Fdn Fiocruz [2]; Casa Oswaldo Cruz Fiocruz [1]; Ctr Informacao Cient & Tecnol [1]; Ctr Pesquisas Gongalo Moniz Fundagao Oswaldo Cruz [1]; Ctr Pesquisas Rene Rachou Fiocruz [1]; Ctr Referencia Helio Fraga [1]; ENSP [1]; Escola Nacl Saude Pbl [1]; Escola Nacl Saude Publ [1]; Escola Nacl Saude Publ Fiocruz [1]; Fdn Oswaldo Cruz [1]; Fdn Oswald Cruz FIOCRUZ [1]; Fdn Oswaldo Cruz MS [1]; FIOCRUZ BH [1]; Fundacao Inst Oswaldo Cruz Fiocruz [1]; Fundacao Oswald Cruz [1]; Fundacao Oswaldo Cruz ILMD FIOCRUZ [1]; Fundaco Oswaldo Cruz FIOCRUZ [1]; IFF FIOCRUZ [1]; Inst Comunicacao & Informacao Cient & Tecnol Saud [1]; Inst Leonidas & Maria Deane [1]; Inst Leonidas & Maria Deane ILMD Fiocruz [1]; Inst Rene Rachou [1]; Natl Sch Publ Hlth [1]; Oswaldo Cruz Fdn ENSP FIOCRUZ [1]; Sergio Arouca Natl Sch Publ Hlth [1]

Quadro 1 (cont.): Filiação do autor na forma padronizada e nas diferentes formas encontradas na base de dados, com o respectivo número de artigos.

Filiação do autor padronizada [n° artigos]	Filiação do autor nas formas encontradas na base [n° artigos]
Univ Fed Minas Gerais [251]	Univ Fed Minas Gerais [236]; UFMG [12]; Fac Odontol UFMG [1]; Terapeuta Ocupac UFMG [1]; Univ Fed Minas Gerias [1]
Univ Estadual Campinas [242]	Univ Estadual Campinas [215]; UNICAMP [17]; State Univ Campinas [3]; Univ Campinas [3]; Univ Estadual Campinas UNICAMP [2]; State Univ Campinas UNICAMP [1]; Univ Campinas Sao Paulo [1]
Univ Fed Pelotas [235]	Univ Fed Pelotas [207]; Fed Univ Pelotas [12]; UFPel [8]; Univ Fed Pelotas UFPel [5]; Univ Fed De Pelotas [3]
Univ Fed Bahia [193]	Univ Fed Bahia [178]; UFBa [11]; Univ Fed Bahia UFBA [2]; ISC UFBA [1]; Univ Fed Bahia Hosp [1]
Univ Fed Sao Paulo [181]	Univ Fed Sao Paulo [163]; UNIFESP [9]; Univ Fed Sao Paulo Unifesp [5]; CPES Unifesp [1]; Fed Univ Sao Paulo UNIFESP [1]; Univ Fed Sao Paulo Unifesp Baixada Santista [1]; Univ Sao Paulo Unifesp [1]
Univ Fed Rio de Janeiro [163]	Univ Fed Rio de Janeiro [134]; UFRJ [23]; Fed Univ Rio De Janeiro [5]; Rio de Janeiro Fed Univ Publ Hlth Nucleus [1]
Univ Fed Rio Grande do Sul [155]	Univ Fed Rio Grande do Sul [104]; Univ Fed Rio Grande Sul [35]; UFRGS [8]; Fundacao Univ Fed Rio Grande [6]; FAMED UFRGS [1]; Fed Univ Rio Grande Sul State [1]
Univ Estado Rio de Janeiro [140]	Univ Estado Rio de Janeiro [126]; UERJ [10]; LCR UERJ [1]; UnATI HUPE UERJ [1]; Univ Estado Rio Janeiro [1]; Univ estadual Rio De Janeiro [1]

Quadro 1 (cont.): Filiação do autor na forma padronizada e nas diferentes formas encontradas na base de dados, com o respectivo número de artigos.

Filiação do autor padronizada [n° artigos]	Filiação do autor nas formas encontradas na base [n° artigos]
Univ Estadual Paulista [130]	Univ Estadual Paulista [65]; UNESP [27]; Univ Estadual Paulista Julio Mesquita Filho [7]; Sao Paulo State Univ [7]; Sao Paulo State Univ UNESP [3]; Univ Estadual Paulista UNESP [3]; UNESP Sao Paulo State Univ [2]; Univ Estadual Sao Paulo [2]; Univ Sao Paulo State [2]; Botucatu Unesp [1]; Dept Psicol Clin UNESP Assis [1]; UNESP Araraquara [1]; Unesp Bauru [1]; UNESP Univ Estadual Paulista [1]; Unidade Estadual Sao Paulo [1]; Univ Estadual Julio Mesquita Filho [1]; Univ Estadual Paulista Julio de Mesquita [1]; Univ Estadual Paulista Julio de Mesquita Filho Un [1]; Univ Estadual Sao Paulo UNESP [1]; Univ Nacl Estadual Sao Paulo [1]; Univ Sao Paulo State Julio Mesquito Filho [1]
Univ Fed Pernambuco [83]	Univ Fed Pernambuco [78]; UFPE [3]; Univ Fed Pernambuco UFPE [2]
Univ Fed Ceara [70]	Univ Fed Ceara [63]; Fed Univ Ceara [2]; Fed Univ Ceara State [1]; LABOMAR UFC [1]; UFC [1]; UFCE [1]; Univ Fed Ceara Bloco Didat [1]
CNEN [67]	IPEN CNEN SP [18]; CNEN [11]; Inst Radioprotecao & Dosimetria [8]; Comissao Nacl Energia Nucl [7]; Comis Nacl Energia Nucl [4]; Inst Radioprotect & Dosimetry [4]; Inst Radiat Protect & Dosimetry [3]; Comis Nacl Energia Atom [2]; Inst Radioprotecao & Dosimetria IRD [2]; IRD CNEN [2]; CNEN SP [1]; Comissao Nacl Energy Nucl [1]; Ctr Reg Ciencias Nucl [1]; Ctr Reg Ciencias Nucl CNEN [1]; Inst Radioprotecao & Dosimetria CNEN [1]; Inst Radioprotecao & Dosimetria IRD CNEN [1]

Quadro 1 (cont.): Filiação do autor na forma padronizada e nas diferentes formas encontradas na base de dados, com o respectivo número de artigos.

Filiação do autor padronizada [n° artigos]	Filiação do autor nas formas encontradas na base [n° artigos]
Secretaria Estado Saude Sao Paulo [56]	Secretaria Estado Saude Sao Paulo [35]; Secretaria Saude Estado Sao Paulo [4]; SES SP [4]; Estado Saude Sao Paulo [2]; Inst Saude Secretaria Estado Saude Sao Paulo [2]; Secretaria Estadual Saude Sao Paulo [2]; CVE SES SP [1]; Sao Paulo State Hlth Dept [1]; Saude Mulher Secretaria Estado Saude Sao Paulo [1]; Secretaria Estado Saude Sao Paulo [1]; Secretar Estado Saude Sao Paulo [1]; Secreteria Estado Saude Sao Paulo [1]; SESSP [1]
Minist Saude [55]	Minist Saude [37]; Minist Hlth [9]; Brazilian Minist Hlth [3]; Ministerio Saude [2]; Minist Jaude [1]; Minist Saude SIA [1]; MS [1]; Secretariat Minist Hlth Brazil [1]
Univ Estadual Londrina [50]	Univ Estadual Londrina [48]; State Univ Londrina [1]; State Univ Londrina UEL [1]
Univ Vale Rio dos Sinos [49]	Univ Vale Rio dos Sinos [43]; Univ Vale do Rio dos Sinos [2]; Univ Vale Rio Sinos [2]; Univ Vale Dos Sinos [1]; Univ Vale Rio Sinos UNSINOS [1]
Univ Fed Espirito Santo [45]	Univ Fed Espirito Santo [42]; Inst Fed Espirito Santo [1]; Univ Espirito Santo [1]; Univ Federal Espirito Santo [1]
Univ Fed Rio Grande do Norte [43]	Univ Fed Rio Grande do Norte [35]; Univ Fed Rio Grande Norte [4]; CB UFRN [1]; UFRN [1]; Univ Fed Rio Grande N [1]; Univ Fed Rio Grande Norte UFRN [1]

Tabela 1: Distribuição do número de artigos antes e depois da padronização do campo “Filiação do primeiro autor”, classificação e percentual de variação para as 20 instituições brasileiras que mais apresentaram artigos na base *Web of Science* relacionados à categoria de assunto “*Public, Environmental & Occupational Health*” no período de 2001 a 2011.

Instituição	Nº de artigos antes da padronização	Classificação antes da padronização	Nº de artigos depois da padronização	Classificação depois da padronização	Percentual de Variação
Univ São Paulo	744	1º	825	1º	10,9
Fiocruz	353	2º	689	2º	95,2
Univ Fed Minas Gerais	236	3º	251	3º	6,4
Univ Estadual Campinas	215	4º	242	4º	12,6
Univ Fed Pelotas	207	5º	235	5º	13,5
Univ Fed Bahia	178	6º	193	6º	8,4
Univ Fed Sao Paulo	163	7º	181	7º	11,0
Univ Fed Rio de Janeiro	134	8º	163	8º	21,6
Univ Fed Rio Grande do Sul	104	10º	155	9º	49,0
Univ Estado Rio de Janeiro	126	9º	140	10º	11,1
Univ Estadual Paulista	65	12º	130	11º	100,0
Univ Fed Pernambuco	78	11º	83	12º	6,4
Univ Fed Ceara	63	13º	70	13º	11,1
CNEN	18	20º	67	14º	272,2
Secretaria Estado Saude Sao Paulo	35	18º	56	15º	60,0
Minist Saúde	37	17º	55	16º	48,6
Univ Estadual Londrina	48	14º	50	17º	4,2
Univ Vale Rio dos Sinos	43	15º	49	18º	14,0
Univ Fed Espirito Santo	42	16º	45	19º	7,1
Univ Fed Rio Grande do Norte	35	19º	43	20º	22,9