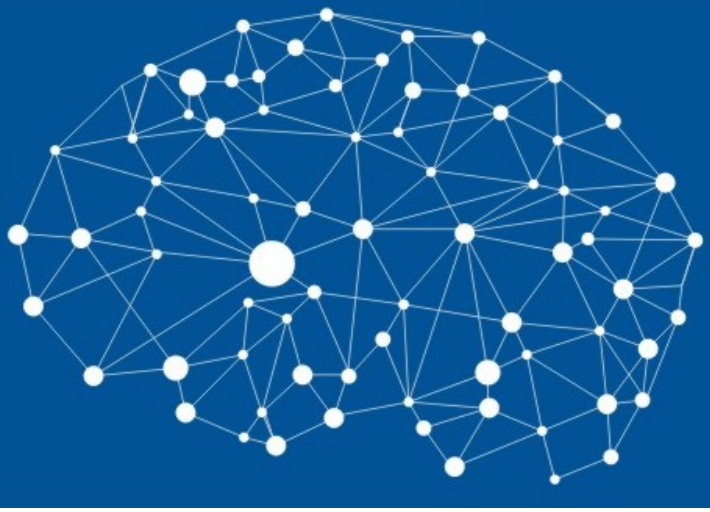

O PROBLEMA DA PADRONIZAÇÃO DAS AFILIAÇÕES DE AUTORES NA BASE DE DADOS WEB OF SCIENCE - O CASO EMBRAPA E SUA SOLUÇÃO

Roberto de Camargo Penteado Filho[1]; Wilson Corrêa da Fonseca Júnior[2]

PENTEADO FILHO, R. C.; JúNIOR, W. C. A. F.. O PROBLEMA DA PADRONIZAÇÃO DAS AFILIAÇÕES DE AUTORES NA BASE DE DADOS WEB OF SCIENCE - O CASO EMBRAPA E SUA SOLUÇÃO In: ENCONTRO BRASILEIRO DE BIBLIOMETRIA E CIENTOMETRIA, 5., 2016, São Paulo. **Anais**... São Paulo: USP, 2016. p. A18

[1, 2] Embrapa

O PROBLEMA DA PADRONIZAÇÃO DAS AFILIAÇÕES DE AUTORES NA BASE DE DADOS WEB OF SCIENCE

O CASO EMBRAPA E SUA SOLUÇÃO

Eixo temático: produção e produtividade científica

Modalidade: apresentação oral

1 INTRODUÇÃO

A produção científica e tecnológica é fundamental no processo de inovação de um país ou instituição científica. No entanto, a avaliação bibliométrica desse desempenho enfrenta há muitos anos um grande desafio: o problema da falta de exatidão das informações registradas em documentos científicos e bases de dados. Este artigo possui como principal objetivo verificar em que medida esse problema ainda persiste na produção científica brasileira, por meio de investigações bibliográfica e bibliométrica. Neste segundo caso, foi realizado, como exemplo, um estudo básico sobre a visibilidade da Empresa Brasileira de Pesquisa Agropecuária – Embrapa na base de dados Web of Science (WoS).

Como se sabe, o problema da falta de exatidão das informações em documentos científicos e bases de dados se deve a diversos fatores, tais como a presença de erros ortográficos ou de impressão, equívocos na classificação de dados e ausência de padronização de informações institucionais, entre outros, conforme registra a literatura especializada (BOURNE, 1977; HOOD, WILSON, 2003; TAŞKIN, AL, 2013). Já na década de 1970 Bourne (1977) chamou a atenção para a frequência e o impacto dos erros de ortografia em bases de dados bibliográficas. Naquela época, ao investigar cerca de 3.600 termos indexados em 11 diferentes bases de dados ele encontrou a presença de erros ortográficos com variação de 23% a menos de 0,5% entre uma base e outra.

Ao analisar o acervo da produção científica da Embrapa com a utilização de diversos softwares (Infotrans, Dataview, Matrisme e MS Excel) Penteado Filho (2007) indicou problemas de erros ortográficos, de catalogação e de digitação, duplicações de registros e presença de autores homônimos. Mais recentemente, ao abordar o problema da padronização das afiliações de autores em índices de citação, Taşkin e Al (2013) analisaram na base de

dados *Web of Science* todos os tipos de publicação (artigos, anais de congressos ou cartas), publicados entre 1928 e 2009, que mencionassem nos campos "endereço" e "país" a Turquia como país de origem. Entre os principais resultados desse trabalho se encontra uma tabela contendo as 20 universidades turcas mais produtivas, com seus respectivos números de publicações, quantidade de erros de indexação e percentual desses erros em relação ao total de publicações. Nesse caso, os percentuais de erros variaram de 12,1% a 0,3% entre as instituições. Subjacente a esses números estavam problemas tais como: erros de caracteres ou de ortografia, erros de indexação, erros de tradução realizados pelos autores, além de problemas de padronização dos endereços das instituições.

Atualmente, entre os problemas que dificultam a correta avaliação da produção científica de um país, instituição ou pesquisador se encontra a ausência de padronização dos nomes de autores e de suas afiliações em bases de dados de ciência e tecnologia (C&T). Em trabalho recente de análise sobre a produção científica da Empresa Brasileira de Pesquisa Agropecuária (Embrapa), Penteado Filho, Fonseca Júnior e Avila (2015) também se empenharam no levantamento dos principais fatores associados a essa questão. Os resultados desse trabalho, realizado na base de dados *Web of Science* (WoS) e apoiado na literatura especializada, se encontram sintetizados neste artigo.

2 METODOLOGIA

Este levantamento reuniu a produção científica da Embrapa nas bases de dados Science Citations Index Expanded (SCI-EXPANDED), Social Science Citation Index (SSCI) e Arts & Humanities Citation Index (A&HCI) da Thomson Scientific - ISI, conhecidas sob a denominação de Web of Science (WoS). Para isso foram considerados todos os registros, em diversas línguas e tipos de documentos (artigos, editoriais, resenhas, entre outros) com as possíveis denominações que fizessem referência à Embrapa no campo "Afiliação de autor" no período compreendido entre 1973 e 17 de julho de 2015.

3 RESULTADOS

A busca teve início pela expressão mais simples. Depois, foram sendo incorporadas outras variações de acordo com as inconsistências encontradas nos artigos referentes ao nome da Embrapa ou de seus centros de pesquisa. O acréscimo dessas variações elevou o número de

registros de 15.956 para 17.794 entre a primeira e terceira etapas. A quarta etapa contemplou uma investigação sobre registros únicos dos centros de pesquisa da Empresa (324 registros), o que aumentou o resultado final para 18.118 registros. A diferença entre esse total e o número de registros da primeira busca (15.956) é de 2.162 registros. Isso significa que, se a busca sobre a produção científica da Embrapa se limitasse aos termos originais adotados na primeira etapa, a sigla/nome da instituição, 11,93% dos documentos não estariam contemplados na sua produção científica. As expressões completas de busca podem ser encontradas no Anexo disponibilizado na Internet da Embrapa (Veja o link para o arquivo no final do trabalho). Veja o quadro geral na Tabela 1.

Tabela 1: Diferentes estratégias de busca do nome da Embrapa na WoS e seus resultados.

ESTRATÉGIAS DE BUSCA DE DOCUMENTOS EMBRAPA NA BASE WOS (17/07/2015)		
ETAPA	**EXPRESSÃO**	**REGISTROS**
1	AD=(Embrapa) OR AD=(EMBRAPA)	15.956
2	AD=(EMBRAPA) OR AD=(EMPRESA BRASILEIRA PESQUISA AGROPECUARIA) OR AD=(Brazilian Org Agr Res) OR AD=(BRAZILIAN ENTERPRISE AGR RES) OR AD=(Brazilian Agr Res Corp) OR AD=(BRAZILIAN AGR RES ENTERPRISE) OR AD=(Brazilian Enterprise Agropecuary) OR AD=(Brazilian Agropecuary Res Corp) OR AD=(BRAZILIAN ORG AGR RES) OR AD=(BRAZILIAN AGR RES CORP) OR AD=(BRAZILIAN ENTERPRISE AGROPECUARY) OR AD=(BRAZILIAN AGROPECUARY RES CORP)	17.438
3	AD=(Einbrapa) OR AD=(EINBRAPA) OR AD=(Embapra) OR AD=(EMBAPRA) OR AD=(Embarapa) AD=(Embraba) OR AD=(EMBRABA) OR AD=(Embraoa) OR AD=(EMBRAOA) OR AD=(Embrape) OR AD=(EMBRAPE) OR AD=(Embrapo) OR AD=(EMBRAPO) OR AD=(Embrara) OR AD=(EMBRARA) OR AD=(Embrpa) OR AD=(EMBRPA) OR AD=(EMBTAPA) OR AD=(Embtapa) OR AD=(Empbrapa) OR AD=(EMPBRAPA) OR AD=(Empera Brasileira Pesquisas Agropecuaria) OR AD=(Empersa) OR AD=(Ambrapa) **etc…**	17.794
4	(“Ctr Nacl Pesquisa & Desenvolvimento Instrumentaca” OR “Ctr Nacl Pesquisa & Gado Corte” OR “Ctr Nacl Pesquisa Agrobiol” OR “Ctr Nacl Pesquisa Algodao” OR “Ctr Nacl Pesquisa Arroz & Feijao” **etc...**) **NOT** (“Ambrapa” OR “AMBRAPA” OR “Brazilian Agr Res Corp” OR “BRAZILIAN AGR RES CORP” OR “BRAZILIAN AGR RES ENTERPRISE” OR “Brazilian Agr Res Enterprise” **etc…**)	324
	RESULTADO GERAL DA BUSCA	17.794 + 324 = **18.118**
	PERCENTUAL DE VARIAÇÕES NÃO CONTEMPLADAS NA BUSCA INICIAL	**(11,93 %)**

Os problemas da falta de exatidão das informações encontrados nessa busca sobre a produção científica da Embrapa se deve a vários fatores, que vão desde a digitação errônea da sigla da Empresa no próprio artigo científico ou no cadastro da base de dados, passando pela ausência de uniformização do nome da instituição em português e outras línguas, ou mesmo pela ausência da sigla e/ou do nome da instituição nesses documentos. De acordo com o levantamento realizado, a origem dos problemas encontrados pode estar na base de dados, no documento científico ou em ambos. Os principais erros encontrados são apresentados a seguir:

a) Erros de digitação de dados ou de redação no próprio documento científico. Exemplos: Embapra / Embarapa / Embraba / EMPRAPA

b) Erros de digitação ou de digitalização de dados do documento científico pela base de dados. Exemplos: Ambrapa / Einbrapa / Embraoa / Empresa brasileira Pesquisa Agropecuciria / Empresa Brasileira Pesquisa Agropecuraria.

c) Erros de digitação e/ou de digitalização de dados no documento científico e na base de dados. Exemplos: Empresa Brasileira de Pesquisas Agropecuária (artigo) / Empresa Brasileira (de) Pesquisas Agropecuária (base de dados)

d) Redução do nome original por extenso da Embrapa pela base de dados. Exemplos: Empresa Brasileira Pesquisas Agr / Brazilian Agr Res Corp

e) Omissão da sigla e/ou nome da instituição no documento científico. Exemplos: Ctr Nacl Pesquisa & Desenvolvimento / Ctr Nacl Pesquisa

f) Prevalência, no documento científico, da sigla e do nome da unidade e/ou departamento em detrimento da sigla e do nome da Embrapa. Exemplo: Laboratório de Análise do Solo, Centro Nacional de... (sigla da unidade)

g) Existência nos documentos científicos de diversidade de nomes em inglês (ou outras línguas) para a Empresa. Exemplos: Brazilian Agricultural Research Enterprise / Brazilian Corporation of Agricultural Research / Brazilian Enterprise for Agricultural Research

h) Coexistência no documento científico da identificação da Empresa em português e inglês (ou outras línguas). Exemplo: Laboratory of bacteriology, Empresa Brasileira de Pesquisa Agropecuária

i) Adoção no documento científico de nomes e abreviações genéricas de unidades ou departamentos da Empresa, em português, inglês e outras línguas, que dificultam a busca em base de dados. Exemplos: Natl Ctr / Natl Res Ctr / Ctr Nacl / Ctr Nacl Pesq

j) Possibilidade de confusão entre as siglas Embrapa (Empresa Brasileira de Pesquisa Agropecuária) e Embrapii (Empresa Brasileira de Pesquisa e Inovação Industrial). Exemplo: Embrap – Empresa Brasileira de Pesquisa e Inovação Industrial – Embrapii / Embrap – Empresa Brasileira de Pesquisa Agropecuária.

3 DISCUSSÃO

Tais problemas não são exclusivos da Embrapa. Eles afetam todas as instituições com as quais a Embrapa produziu documentos científicos. Veja, na Tabela 2, um resumo da diversidade de grafias referentes aos nomes dos cinco principais parceiros da Embrapa na elaboração dos seus artigos. Para efeito didático, de demonstrar a extensão e alcance deste problema de padronização, acrescentamos cada uma dessas listas no Anexo disponibilizado em link da Internet da Embrapa.

Tabela 2. Estatísticas de artigos e grafias dos cinco principais parceiros da Embrapa em artigos na WoS.

Class.	Instituição	Total Artigos	# Afiliações	%
1	UFV	1125	656	58
2	UNESP	1008	1223*	121
3	UnB-BR	916	680	74
4	USP-ESALQ	870	896*	103
5	UFLA	678	473	69

* Este número pode ser explicado pela existência de dois ou mais autores da UNESP e da USP-ESALQ num mesmo artigo que citaram sua afiliação de maneira diferente.

Quando examinadas no nivel micro, isto é, do centro de pesquisa, esses erros de padronização têm o poder de multiplicar-se de forma exponencial. Por exemplo, a Embrapa

Recursos Genéticos e Biotecnologia possui 1.545 artigos científicos publicados no período na WoS nos quais foram encontrados 1.188 diferentes afiliações.

Para efeito de contagem da produção científica na base WoS, cada uma das diferentes grafias (1.188) corresponde a uma instituição. Essa é a principal razão porque, apesar de produzir artigos suficientes para figurar sozinha entre as 50 primeiras instituições do país, a Embrapa Recursos Genéticos e Biotecnologia aparece, na base não tratada, na sua primeira menção, "Embrapa Recursos Genet & Biotecnol, BR-70770900 Brasilia, DF, Brazil", com 174 artigos. Esse score corresponderia aproximadamente ao 700º lugar das instituições brasileiras. Essas 1.188 diferentes entradas da Embrapa Recursos Genéticos e Biotecnologia também estão acessíveis no Anexo deste artigo.

Esse problema se reproduz, por exemplo, numa universidade, quando se desce ao nível do Departamento e vai além, nos dois casos, da Embrapa e de Instituições de Ensino Superior (IES), quando se contempla mais um nível, ou seja, ao de Laboratório ou Grupo de Pesquisa. Nesse nível ocorrem os piores erros que, muitas vezes, impedem inclusive a atribuição do artigo a qualquer instituição. É comum a inversão da afiliação, citando primeiro o laboratório, em seguida o departamento e, por último, a instituição. O bom senso indica a ordem inversa, sempre; instituição, departamento, laboratório.

A inversão de afiliação carrega consigo o pior erro de exatidão, que é a elisão da instituição. Neste caso aparecem entradas como "Dept Geotechnol, BR-13405970 Piracicaba, SP, Brazil", "Dept Biol Celular, Brasilia, DF, Brazil", "Dept Tecnol Agroind & SocioeconomicoRural, Araras, SP, Brazil", "Depto Desenvolvimento Ensino, Sao Paulo, Brazil" ou "Dept Cientifico, Araraquara, SP, Brazil". Este erro é comum também em instituições estrangeiras: "Lab Cytogenet & Gebine Res, B-3000 Louvain, Belgium", "Lab Invest Aplicada, Cordoba 14080, Spain" ou "LISC, Clermont Ferrand, France". Nesses casos há uma grande probabilidade de que a afiliação desse artigo será perdida na contagem em rankings internacionais.

A recuperação de 18.188 artigos da Embrapa entre 1973 e 2015 a credencia como uma das dez primeiras instituições produtoras de artigos científicos indexados na WoS. No entanto, a maioria dos rankings de instituições brasileiras realizada a partir dessa base consegue enxergar melhor as universidades, ao passo que a Embrapa é vista de forma parcial

(BRASIL, 2008; GOIS, 2008; GREGOLIN et al., 2005; LETA; CRUZ, 2003). Trata-se de uma perda considerável de visibilidade pública na base WoS, tanto da Empresa como de seus centros de pesquisa, parcialmente corrigida internamente pelo trabalho de acompanhamento da produção científica da Embrapa na WoS realizada pela Secretaria de Gestão e Desenvolvimento Institucional – SGI, vinculada à própria Embrapa.

No caso geral das instituições brasileiras esses erros repetidos e constantes acarretam a perda de lugares preciosos nos rankings de IES, que estão se tornando cada vez mais populares em todos os continentes.

4 RECOMENDAÇÕES E CONCLUSÕES

A partir da constatação desses problemas os resultados deste trabalho foram apresentados internamente a um grupo de trabalho coordenado pela Embrapa Informação Tecnológica, que propôs à presidência da Embrapa a regulamentação da afiliação institucional dos empregados da Empresa em publicações nacionais e internacionais. Essa proposta se materializou com a publicação de uma resolução normativa interna em março de 2016 com as seguintes determinações:

1. Em todas as publicações nacionais e internacionais, a afiliação institucional do autor deve ser indicada pela assinatura síntese da unidade. Exemplos: Embrapa Gado de Corte; Embrapa, Departamento de Pesquisa e Desenvolvimento – DPD;

2. Os nomes das unidades centrais e descentralizadas devem ser mantidos sem tradução, em todas as publicações nacionais e internacionais;

3. O endereço institucional não deve ser traduzido, devendo permanecer em português.

4. No endereço institucional, apenas o nome da unidade central ou descentralizada deve ser utilizado, sem a indicação de qualquer laboratório ou área, mesmo em artigos em coautoria com universidades.

5. A Embrapa Informação Tecnológica irá inserir essas orientações no Manual de Editoração da Embrapa e orientará as unidades no cumprimento dessa determinação.

6. Os Comitês Locais de Publicação (CLPs) devem garantir o cumprimento dessa norma.

Este trabalho e seu Anexo visam chamar a atenção de todos os responsáveis e dirigentes de instituições brasileiras para a extensão do problema.

REFERÊNCIAS

BOURNE, C.P. Frequency and impact of spelling errors in bibliographic databases. **Information Processing & Management**, Vol.13, n.1, p.1-12, 1977;

BRASIL. Ministério da Ciência e Tecnologia. **Indicadores Nacionais de Ciência & Tecnologia.** Disponível em: <http://www.mct.gov.br>. Acesso em: 11 dez. 2008.

EMBRAPA. Resolução Normativa n°4. **Boletim de Comunicações Administrativas**. Embrapa: Brasília, 2016, p.8-9.

GOIS, A. ITA lidera em produtividade científica. **Folha de S. Paulo**, São Paulo, 14 jan. 2008. Caderno Ciência, A10.

GREGOLIN, J. A. R.; HOFFMANN, W. A. M.; FARIA, L. I. L.; QUONIAM, L.; QUEYRAS, J. Análise da produção científica a partir de indicadores bibliográfi cos. In: LANDI, F. R.; GUSMÃO, R. (Coord.). **Indicadores de ciência, tecnologia e inovação em São Paulo 2004.** São Paulo: FAPESP, 2005. 2v., 992 p. Disponível em: <http://www.fapesp.br/indicadores>. Acesso em: 27 jul. 2005.

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

LETA, J.; CRUZ, C. H. de B. A produção científica brasileira. In: Viotti, Eduardo. B.; Macedo, Mariano de M. (Orgs.). **Indicadores de ciência, tecnologia e inovação no Brasil.** Campinas: Editora da Unicamp, 2003, 615p.

PENTEADO FILHO, **R. de C. Création de systèmes d'intelligence dans une organisation de recherche et développement avec la scientométrie et la médiamétrie.** Tese (Doutorado). Université du Sud, Toulon Var - Toulon_FR, setembro 2006, 328p.

PENTEADO FILHO, R. de C.; FONSECA JÚNIOR, W. C. da; AVILA, A. F. D. Perfil da Produção Científica da Embrapa entre 2004 e 2013: Oportunidades e Desafios. Documentos (Embrapa SGI) (1679-4680), v. 17, 2015, no prelo.

TAŞKIN, Z.; AL, U.. Standardization problem of author affiliations in citation indexes. **Scientometrics**, Vol.98, n.1, p.347-368, 2013.

ANEXO disponível na Internet no link:
http://www22.sede.embrapa.br/web/sge01/estatisticaagricola/anexopadronizacao5ebbc.pdf