Normalização de nomes de autores em fontes de informação institucionais: proposta de um método automático de verificação de erros

Rogério Mugnaini
Luciano Antonio Digiampietri
Laucivaldo Cardoso de Oliveira
Sueli Mara Soares Pinto Ferreira

RESUMO

A recuperação da produção científica por autoria é um desafio para diversos mantenedores de bases de dados, devido à ambiguidade causada por problemas derivados da falta de controle no momento da indexação. Este estudo apresenta um método automático para verificação de ocorrência de erros em dados de autorias da base de produção científica da Universidade de São Paulo (Dedalus) tomando como base o banco institucional de recursos humanos. Por meio de algoritmos de busca aproximada, compararam-se esses dados com os dados de autoria registrados no banco de produção científica de quatro unidades da USP (período de 2006-2010). Com base nesse estudo piloto foi possível estabelecer mecanismos de interoperabilidade entre o banco da produção e os bancos institucionais de recursos humanos, além de permitir mapear a porcentagem de erros, desenvolver mecanismos de interferência e estabelecer um cronograma para ampliar o estudo as demais unidades USP, bem como procedimentos de normalização.

PALAVRAS-CHAVE: Produção científica. Autoridade de autor. Normalização. Processamento automático. Indexação.

1 Introdução

O sistema de Ciência e Tecnologia de toda e qualquer instituição ou país passa pela avaliação de sua produção científica. Produção esta que necessita estar devidamente indexada em bancos de dados normalizados favorecendo sua apresentação e classificação de distintas maneiras e a partir de diversificados olhares e indicadores. Dentre as diversas dificuldades e problemas inerentes ao processo de indexação vivenciado pelos inúmeros bancos de produção científica existentes no mundo, um dos mais complexos se refere especificamente a indexação de autoria dos trabalhos.

Tal problema se origina no próprio documento quando o autor se autodenomina de distintas maneiras em diferentes momentos (nome por extenso, abreviado, de casado, fantasia ou outras derivações). Essa falta de normalização gera um encadeamento de problemas de ambiguidade nas bases de produção científica, tendo em vista que o indexador normalmente vai entrar os nomes dos autores de acordo com o documento que tem em mãos. Com o crescimento do banco de produção esse problema se avoluma e interfere diretamente na melhor e mais acurada análise de produtividade de uma dada instituição ou país.

Esse desafio tem sido encarado por pesquisadores de diversas áreas, com uma maior prevalência na Ciência da Computação e Ciência da Informação.

A Ciência da Computação lança mão de métodos complexos, que podem ser subdivididos entre **supervisionados** (TORVIK et al., 2005; FERREIRA et al., 2010) ou **não supervisionados** (HAN et al., 2005; CARVALHO et al., 2011). **Métodos supervisionados** são aqueles que dispõem de um conjunto de dados previamente identificados por usuários e que utilizam de técnicas de aprendizagem para criar as regras que irão classificar novos dados. **Métodos não supervisionados** são aqueles que não dispõem deste tipo de informação e tentam agrupar os dados de acordo com alguns critérios, por exemplo, agrupar nomes com grafias similares. Por fim, cada grupo de nomes similares é associado a um autor. Nota-se que a área lida com grandes quantidades de informação, não se limitando às bases de dados de produção científica, mas usando fontes diversas da Web, como o Google (MANN; YAROWSKY, 2003) ou a Wikipedia (CUCERZAN, 2007).

Já a área de Ciência da Informação, normalmente restringe suas análises às bases de dados de produção científica, e chega a lançar mão de algoritmos computacionais (COSTAS; BORDONS, 2007), mas também utiliza métodos mais simples, por exemplo, fazendo uso do contexto – lista de autores e departamen-

to da universidade – para normalização dos nomes dos autores na produção científica de uma universidade (GUERRERO-BOTE et al., 2002). Um dos instrumentos utilizados por grandes bibliotecas é o banco de autoridade de autor. No entanto, trata-se de uma ferramenta que exige grande disciplina dos profissionais envolvidos, os quais deverão, sempre e impreterivelmente, proceder a uma busca do nome normalizado dos autores, antes de iniciar a indexação.

A problemática da ambiguidade de nomes de autores consiste, de modo geral, de duas situações:

a) a homonímia, que exige que se tenha que distinguir entre diferentes pessoas com um mesmo nome e
b) as diversas variações de nome de uma mesma pessoa.

Uma vez identificados os homônimos ou nomes similares, deve-se buscar outras informações que permitam identificá-los (como uma mesma pessoa) ou diferenciá-los, o que normalmente se faz utilizando diversas informações, que Kang e outros (2009) subdividem em:

a) características bibliográficas domínio-independentes, como data ou lugar de nascimento, e-mail ou endereço postal, e
b) características contextuais domínio-específicas, como co-autores e título do artigo, no caso de dados bibliográficos.

A utilização de nomes de coautores, para desambiguidade de homônimos, vem repercutindo também em estudos recentes de análise de redes sociais (ou redes de coautoria) (KANG et al., 2009; SHIN et al., 2010; MÉNDEZ-VÁSQUEZ et al., 2012).

Uma das bases de dados internacionais mais popular mundialmente, a *Web of Science*, frente a esse problema, criou mecanismos visando oferecer ao usuário instrumento que lhe permite descobrir se os resultados condizem ou não com o autor buscado. Para tanto, oferece a opção de seleção por área temática de publicação ou ainda afiliação.

No Brasil, essa problemática de falta de normalização também está presente em diversas bases, como a Plataforma Lattes[1] frente às diferentes maneiras que os distintos autores e coautores cadastram uma mesma publicação em seus currículos (ALCÁZAR et al, 2011; DIGIAMPIETRI ; SILVA, 2011).

Frente à complexidade e internacionalidade desse problema, surgiu recentemente o projeto intitulado *Researcher and Contributor ID (Orcid)*, que busca apresentar uma solução para problemas de ambiguidade na identificação dos autores de todo o mundo. Lançado oficialmente em agosto de 2010, caracteriza-se por ser uma organização sem fins lucrativos, composta por entidades influentes como universidades, institutos de pesquisa, agências de fomento e editores de revistas científicas. Propõe um sistema

[1] Plataforma Lattes: <http://lattes.cnpq.br/> acessada em 12 de maio de 2012.

global gratuito, aberto (no acesso à informação e no uso de *Open Source Software*) e inclusivo visando à geração de um identificador único e universal aos pesquisadores do mundo inteiro. Ao se cadastrar e gerar seu próprio identificador, os autores poderão ainda incluir e compartilhar informações com seus perfis acadêmicos, trabalhos elaborados, além de confirmar publicamente se um item realmente é de sua autoria (GARCÍA-GÓMEZ, 2012 ; BILDER, 2011 ; ORCID, 2012).

O projeto definiu que o identificador universal de autor deverá ser claro, permanente e unívoco. Por enquanto, o Orcid apresentará um número com 16 dígitos compatíveis com a norma ISO (ISO 27729), expressos como *Uniform Resource Locator* (URL) antecedido de http://orcid.org/, onde, a cada quatro dígitos, será inserido um hifen que visa facilitar a leitura. Para exemplificar, os números Orcid ficarão com a seguinte configuração: http://orcid.org/0000-0002-8205-121X, sendo que a adesão ocorrerá através de cadastro no próprio site do projeto ou nas *Application Programming Interface* (APIs) de *sites* colaboradores. A proposta do projeto é entrar em vigor em outubro de 2012 (ORCID, 2012).

Desse modo, verifica-se que o tema de normalização de autoria em produção científica é um importante campo bibliográfico para estudo de produtividade, em especial para instituições de ensino e pesquisa. Nesse contexto, a Universidade de São Paulo (USP) foi pioneira no país quando, em agosto de 1990, lança a Resolução 3716 estabelecendo normas de coleta das informações bibliográficas referentes à produção intelectual (científica, técnica, acadêmica e artística) gerada internamente, a fim de centralizá-las, armazená-las e tratá-las tecnicamente, visando facilitar sua utilização pela comunidade. Desde então, o Sistema Integrado de Bibliotecas da USP (SIBiUSP) já indexou mais de 540 mil itens de produção.

No entanto, ainda atenta à necessidade de definir e cruzar diversos indicadores tanto bibliográficos, administrativos como acadêmicos, sempre se pautando em mecanismos dinâmicos e sistêmicos, a Universidade criou o Grupo Permanente de Integração de Dados do Sistema Acadêmico da USP[2], vinculado diretamente a Reitoria. Tal grupo tem a finalidade de integrar informações demográficas, de desempenho e de financiamento nas áreas de atividades-fim da Universidade, disponíveis nos diferentes sistemas e bases de dados, mantendo um conjunto de dados consolidado e continuamente atualizado, bem como expandir essa capacidade de informação para fins de planejamento, gestão e comunicação externa.

Frente ao contexto institucional mencionado, este estudo apresenta um método automático para verificação de ocorrência de erros em dados de autorias da base de produção científica da

[2] Portaria GR Nº 5075, de 25 de maio de 2011.

Universidade de São Paulo (Dedalus). O método desenvolvido é **não supervisionado**, pois combina informações de contexto (lista de autoridades) com algoritmos de busca aproximada de nomes.

2 Metodologia

A descrição da metodologia empregada está estruturada em duas partes:

a) descrição das fontes de informação utilizadas e
b) tratamento dos dados desde sua obtenção, organização e processamento automático.

2.1 Fontes de Informação

Neste trabalho foram utilizadas duas bases de dados. A primeira contém a descrição da produção científica dos docentes da USP enquanto a segunda contém dados coorporativos de todos os funcionários desta Universidade. Estas bases são descritas nas subseções a seguir.

2.1.1 Dedalus

Base referencial da produção intelectual da USP, coordenada pelo SIBi/USP, com processo de alimentação e indexação sob responsabilidade das equipes distribuídas em sessenta e oito bibliotecas dispersas por dez cidades do Estado de São Paulo e cobrindo todas as áreas do conhecimento. A descrição física dos documentos segue as regras do *Anglo-American Cataloguing Rules 2* (*AACR2*) e do formato *Machine Readable Cataloging* (*MARC*).

Até esse momento, a produção intelectual da USP está classificada em 44 tipos de documentos, de acordo com a característica do item a ser catalogado, a saber: artigo de periódico, curadoria, parecer técnico, monografia/livro, parte de material didático, patente, programa de computador, relatório técnico, *website* dentre muitos outros. Os registros bibliográficos no formato MARC apresentam pontos de acesso que podem ser identificados por um dos campos de autoria principal (100, 110, 111 e 130) e/ou de entrada secundária (700, 710,711 e 730), respeitando as regras do *AACR2*. O campo denominado 100 **Entrada principal--nome pessoal** apresenta como atributo o nome do responsável pelo conteúdo intelectual da obra. Existem obras que apresentam até três autores e, segundo as regras do *AACR2*, deve ser criada uma entrada no campo 100 e outras duas no campo 700 (**Entrada secundária--nome pessoal**) do formato *MARC*, sendo o último destinado ao acesso à entrada secundária de autoria pessoal. Já o campo 946, Campo local para informações USP, é utilizado para citação de informações referentes aos Docentes, Servidores Técnico-administrativos, mestrandos e doutorandos vinculados à Universidade.

Além do módulo de indexação, o Sistema Dedalus continha, até 2009, um banco de autoridades de autor para registro normalizado dos nomes de autores USP. No entanto, na migração do sistema ALEPH versão 300 para ALEPH versão 500, em 2010, tal banco sofreu alguns danos.

2.1.2 Tycho - Sistema de apoio à avaliação e a gestão institucional da USP

Deparando-se com certa dificuldade para elaboração de relatórios gerenciais holísticos e "[...] extração *ad-hoc* de dados da base de dados dos sistemas corporativos [...]" (UNIVERSIDADE..., 2012), a USP desenvolveu o Sistema Tycho. Este sistema coleta e integra dados da base de dados corporativos existentes nos sistemas centrais mantidos pelo Departamento de Informática da Reitoria (DI) da USP e do Currículo Lattes e Grupos de Pesquisa, mantidos pelo CNPq. Com base nessa agregação de dados de diferentes fontes, o sistema gera grafos de colaboração e indicadores da produção intelectual no período compreendido entre 1996-2012. A estrutura de suporte desse sistema é o Banco Institucional de Recursos Humanos USP de onde se obtém o nome completo dos autores, dados de vínculo institucional e documentos de identidade.

2.2 Tratamento dos dados

O tratamento de dados executado neste trabalho foi dividido em duas tarefas principais: a obtenção e organização dos dados e o processamento automático dos mesmos. Estas tarefas são descritas nas subseções a seguir.

2.2.1 Obtenção e organização

O Sistema Tycho foi utilizado para obtenção do nome completo dos docentes em atividade na USP. Para isso duas ferramentas diferentes foram desenvolvidas. A primeira foi utilizada para encontrar e organizar os *sites* correspondentes às 51 unidades da USP (segundo a definição utilizada no próprio Sistema Tycho). Com esta informação, uma nova ferramenta foi desenvolvida para recuperar o nome completo dos docentes de cada uma das unidades. Os nomes extraídos do Tycho foram considerados os nomes corretos (e completos) a serem utilizados como referência nas próximas etapas de processamento e análise dos dados. Este processo identificou 5.785 docentes ativos na USP.

A base de dados Dedalus foi utilizada para obtenção dos registros bibliográficos da produção intelectual em formato *MARC*, facilitando assim a automatização da organização e processamento dos dados. O período selecionado para estudo

corresponde a 2006-2010, limitando-se a produção de apenas quatro unidades da USP: Escola de Artes, Ciências e Humanidades (EACH), Escola de Comunicações e Artes (ECA), Faculdade de Educação (FE) e Instituto de Física de São Carlos (IFSC), totalizando 12.628 registros bibliográficos. Tomaram-se, para cada publicação, tanto a lista de todos os autores como a lista de autores da USP.

2.2.2 Processamento Automático

Um algoritmo foi desenvolvido para o processamento automático dos dados com dois propósitos principais:

a) identificar as diferentes formas que o nome de um docente aparece nos registros bibliográficos do Dedalus comparando-se com o banco de nome certo extraído do Tycho, e
b) verificar a existência de algum tipo de inconsistência nos dados analisados.

Antes de efetuar o processo de comparação entre os dados das duas bases (Dedalus e Tycho), todos os acentos e cedilhas dos registros foram removidos e outros sinais gráficos como apóstrofes ou hífens, sendo substituídos por espaços em branco. Isto foi feito porque a maioria dos registros está cadastrada dessa forma e, portanto, a utilização dos mesmos poderia reduzir a eficiência do algoritmo.

Visando partir de uma busca mais precisa para uma busca aproximada, foram utilizadas as seguintes estratégias:

a) busca do nome completo do autor dentro dos registros do Dedalus, exatamente da maneira que ele aparece no sistema Tycho;
b) busca pelo nome do docente, permitindo-se que um ou mais nomes do meio estejam abreviados. Exige-se assim que, ao menos, o primeiro e o último nome estejam completos e que todos os demais nomes estejam completos ou abreviados (mas nenhum nome ou abreviação pode estar faltando);
c) considerou-se os mesmos critérios da segunda, porém permitindo a ausência ou excesso do último sobrenome (e neste caso, exigia-se que o sobrenome anterior fosse encontrado). Esta estratégia foi desenvolvida para tratar principalmente dois casos: o caso das pessoas que adotam um novo sobrenome após o casamento e nomes que são encerrados por "Filho", "Júnior" e outros do gênero;
d) mais complexa do que as estratégias anteriores, esta permite uma combinação de várias diferenças entre os nomes que estão sendo comparados. Especificamente, permite-se

que o primeiro nome esteja abreviado, que haja um nome ou uma abreviação sobrando/faltando em um dos nomes buscados, que nomes não abreviados sejam considerados compatíveis caso as diferenças entre eles sejam de no máximo duas letras (utilizando-se um algoritmo de distância de edição para calcular estas diferenças). Devido a grande combinação de situações possíveis na comparação entre dois nomes, a estratégia adotada utilizou um esquema de pontuações positivas e negativas para cada situação. Por exemplo, se os dois nomes possuíssem o último sobrenome em comum, seria atribuída uma nota positiva. Por outro lado, se houvesse uma pequena diferença entre os sobrenomes (até duas letras de diferença) seria atribuída uma nota positiva, porém menor que a primeira. O mesmo princípio de pontuação é utilizado para as diferentes situações: nomes ausentes ou em excesso, abreviações, etc. Se a pontuação final, após a comparação de todas as partes dos dois nomes completos em verificação for positiva, então o sistema consideraria que conseguiu encontrar o nome buscado, caso contrário consideraria que o nome não havia sido encontrado. Os valores das pontuações utilizados para cada situação foram estabelecidos de maneira empírica.

O Quadro 1 apresenta alguns exemplos de nomes que foram corretamente identificados utilizando-se cada uma das estratégias de busca apresentadas. Observa-se, nos últimos dois casos, a presença de erros de digitação.

Quadro 1 - Exemplos de nomes identificados utilizando cada estratégia de busca

ESTRATÉGIA DE BUSCA	NOME PRESENTE NO TYCHO	NOME ENCONTRADO NO DEDALUS
a)	André Felipe Simões	ANDRE OLMOS SIMOES
	Candido Ferreira Xavier de Mendonça Neto	CANDIDO FERREIRA XAVIER MENDONCA NETO
	Carlos de Brito Pereira	CARLOS DE BRITTO PEREIRA
b)	Carmen Lucia Cardoso	CARMEN L. CARDOSO
	Ciro Juvenal Rodrigues Marcondes Filho	CIRO J. R. MARCONDES FILHO
	Carlos Alberto de Moura Ribeiro Zeron	CARLOS A. M. R. ZERON
	Elizabeth Ferreira Cardoso Ribeiro Azevedo	ELIZABETH F. C. R. AZEVEDO
c)	Osvaldo Novais de Oliveira Junior	OSVALDO NOVAIS DE OLIVEIRA
	Oswaldo Baffa Filho	OSWALDO BAFFA
	Ângela Maria Machado de Lima Hutchison	ANGELA MARIA MACHADO DE LIMA
d)	Thereza Christina Vessoni Penna	T. C VESSONI PENNA
	Thomás Augusto Santoro Haddad	T HADDAD
	Andréa Simone Stucchi de Camargo Alvarez Bernardez	ANDREA SIMONE STUCCHI DE CAMARGO
	Claudia Elisabeth Munte	CLAUDIA ELIZABETH MUNTE
	Carlos de Brito Pereira	CARLOS DE BRITTO PEREIRA

Fonte: Os autores

O Quadro 2 apresenta alguns nomes que o sistema de busca corretamente identificou como diferentes por mais que estes nomes possuíssem algumas palavras iguais ou parecidas.

Quadro 2 – Exemplos de nomes corretamente classificados como diferentes pelo sistema

NOME PRESENTE NO TYCHO	NOME ENCONTRADO NO DEDALUS
Maria Cristina Monteiro de Souza Gugelmin	MARIA C. C. C. SOUZA
Alessandra Lopes de Oliveira	ALECSANDRA MATIAS DE OLIVEIRA
Marcia Regina da Silva	MARINA VIEIRA DA SILVA
Maria Cristina Castilho Costa	CRISTINA COSTA
Alyne Simões Gonçalves	ALINE LIMA GONCALVES

Fonte: Os autores

Após identificar estas correspondências entre os dados do Dedalus, a mesma função de resolução de entidades foi utilizada para identificar a correspondência entre os nomes de docentes USP encontrados no Dedalus e os nomes encontrados no sistema Tycho. Para cada nome presente no sistema Tycho, das quatro unidades da USP avaliadas, foi criada uma lista contendo as diferentes maneiras que o docente foi citado no Dedalus e instalado um contador para somar quantas vezes, cada uma das maneiras identificadas, havia sido utilizada.

Por fim, uma função foi desenvolvida para identificar e contabilizar quais foram as principais diferenças que ocorreram nas citações dos docentes. Elas foram classificadas em oito categorias:

a) nomes a menos nas citações;
b) sobrenomes a menos nas citações;
c) abreviações;
d) nomes com diferenças (com palavras diferentes);
e) nomes parecidos (erro de digitação);
f) sobrenomes parecidos (erro de digitação);
g) sobrenomes a mais, e
h) nomes invertidos (ordem invertida).

O Quadro 3 apresenta um exemplo de cada um dos tipos de diferenças que ocorreram nas citações dos docentes. Em particular, os dados de nomes com diferenças são aqueles que irão ser explorados nos trabalhos futuros para a confirmação se pertencem ou não a mesma pessoa.

Quadro 3 – Exemplos de tipos de diferenças encontradas nas citações

ESTRATÉGIA DE BUSCA UTILIZADA	NOME PRESENTE NO TYCHO	NOME ENCONTRADO NO DEDALUS
Nomes a menos	Alvaro Augusto Comin	ALVARO COMIN
Sobrenomes a menos	Andrea Viude Castanho	ANDREA VIUDE
Abreviações	Terezinha Fátima Tagé Dias Fernandes	TEREZINHA F. T. D. FERNANDES
Nomes com diferenças	Roberto da Silva	ROBERTO LEAL LOBO E SILVA FILHO
Nomes parecidos (erro de digitação)	José Sergio Fonseca de Carvalho	JOSE SERGIO DA FONSECA CARVALHO
Sobrenomes parecidos (erro de digitação)	Ajith Kumar Sankarankutty	AJITH KUMAR SANKARANKUTY
Sobrenomes a mais	Dania Emi Hamassaki	DANIA EMI HAMASSAKI BRITTO
Nomes invertidos (ordem invertida)	*Nenhum caso foi encontrado.*	*Nenhum caso foi encontrado.*

Fonte: Os autores

3 Análise dos dados coletados e resultados

Do total de 12.628 registros bibliográficos da produção científica, das quatro unidades, pode-se observar sua distribuição entre as mesmas, considerando-se os diversos tipos de documento. O Quadro 4 apresenta a porcentagem dos registros considerando o número total de autores USP em relação ao número total de autores por artigo. É possível notar que 0,1% dos registros (oito registros) possuem dois autores USP, mas o total de autores é apenas um, o que indica um erro no cadastramento.

Quadro 4 – Porcentagem dos registros em relação ao número de autores total e número de autores USP

		Número de autores			
		1	2	3	4 ou mais
Número de autores USP	1	37,1%	16,1%	6,8%	14,2%
	2	0,1%	4,2%	3,7%	11,4%
	3	0,0%	0,0%	1,3%	5,1%
	4 ou mais	0,0%	0,0%	0,0%	3,7%

Fonte: Os autores

.

As produções apresentadas na Tabela 1 possuem, ao todo, 1.137 autores docentes diferentes, destes, pode-se mapear 74,2% diretamente no Tycho, o que revela que os restantes 25,8% (ou 293) dos nomes podem precisar de normalização. Por outro lado, quando se considera o número total de ocorrências destes nomes entre todos os autores que participam da produção das quatro unidades, a porcentagem de nomes passíveis de normalização diminui para 7,7% (de um total de 28.284).

Tabela 1 – Distribuição da produção segundo tipo de material e unidade da USP

Tipo de material	EACH	ECA	FE	IFSC	Total
Trabalho de Evento-Resumo	189	2	208	3.497	**3.896**
Artigo de Periódico	654	366	580	1.408	**3.008**
Parte de Monografia/Livro	156	486	628	56	**1.326**
Artigo de Jornal-Dep/Entr	5	47	589	179	**820**
Trabalho de Evento	216	78	197	198	**689**
Monografia/Livro-Ed/Org	26	67	422	15	**530**
Monografia/Livro	41	165	170	7	**383**
Artigo de Jornal	6	118	104	58	**286**
Parte de Monografia/Livro-Apres/Pref/Posf	15	25	218	6	**264**
Artigo de Periódico-Dep/Entr	9	14	161	64	**248**
Editor de Periódico	3	2	161	55	**221**
Trabalho de Evento-Anais Periódico	17	2	14	127	**160**
Trabalho de Evento-Resumo Periódico	46	0	1	89	**136**
Outros	134	154	248	125	**661**
Total	**1517**	**1526**	**3701**	**5884**	**12.628**

Fonte: Os autores

O Quadro 5 contém os diferentes tipos de produção cadastrados no Dedalus para cada uma das unidades da USP analisadas

Com relação aos tipos de variação encontrados, podem-se observar, no Quadro 5, que os nomes incompletos são os mais recorrentes, afetando um total de 1.466 ocorrências. Por outro lado, ao se comparar a média de ocorrências por docente, destacam-se nomes incompletos (9,2), devido à omissão de sobrenomes. Alguns nomes chegam a apresentar até quatro variações, conforme observado no Quadro 6, por outro lado, a maior parte dos nomes passíveis de normalização tem apenas uma variante.

Quadro 5 – Ocorrências dos diversos tipos de variação da escrita dos nomes a normalizar

Tipos de variações de nomes encontradas	Total de Docentes	Total de ocorrências	Média de ocorrências por docente
Nomes a menos	185	1466	*7.9*
Sobrenomes a menos	41	376	*9.2*
Abreviações	79	179	*2.3*
Nomes com diferenças	15	56	*3.7*
Sobrenomes parecidos (erro de digitação)	9	43	*4.8*
Sobrenomes a mais	7	37	*5.3*
Nomes parecidos (erro de digitação)	13	35	*2.7*
Nomes invertidos (ordem invertida)	0	0	-

Fonte: Os autores

Quadro 6 – Ocorrências das quantidades de variação de um mesmo nome

Quantidade de variações do nome	Total de Docentes		Total de ocorrências	
	Freq.	%	Freq.	%
1	249	*84.7*	7935	*63.0*
2	34	*11.6*	2632	*20.9*
3	8	*2.7*	1934	*15.4*
4	3	*1.0*	96	*0.8*
5 ou mais	-	-	-	-

Fonte: Os autores

Quadro 7 – Exemplos alguns nomes com quatro variações em suas citações

Nome de autores no Tycho	Ocorrências do nome completo nas citações	Variações do nome	Ocorrências de cada variação
Elizabeth Ferreira Cardoso Ribeiro Azevedo	12	ELIZABETH F. C. R. AZEVEDO	4
		ELIZABETH R AZEVEDO	1
		ELIZABETH R. AZEVEDO	1
		ELIZABETH RIBEIRO AZEVEDO	2
Orlando de Castro e Silva Júnior	61	O. CASTRO E SILVA	1
		ORLANDO CASTRO E SILVA JUNIOR	1
		ORLANDO DE CASTRO E SILVA	3
		ORLANDO DE CASTRO SILVA JUNIOR	2
Vivian Grace Fernandez Davila Urquidi	3	VIVIAN DAVILA URQUIDI	1
		VIVIAN G. F. D. URQUIDI	2
		VIVIAN G. F. DAVILA URQUIDI	1
		VIVIAN URQUIDI	1

Fonte: Os autores

4 Considerações finais

Este estudo piloto apresentou uma análise de parte dos registros da produção intelectual da USP, indexada no sistema Dedalus da USP, focando-se no padrão e norma de registro de autoria. Para tanto, foi necessária a identificação de uma base de referência para os nomes dos autores. Neste caso, foi utilizada a base Tycho que contém os dados da base de recursos humanos do corpo docente, discente e funcional da Universidade. Como estratégia de trabalho foram desenvolvidas ferramentas para a identificação e contagem automática das variações dos nomes, bem como para identificar potenciais problemas nos registros bibliográficos.

Como resultado verificou-se que, embora não expressivo, existe um percentual de registros de autoria com problemas. Tal situação favorece a perda de qualidade ou mesmo resultados distorcidos quando tais dados forem utilizados para gerar indicadores de produtividade ou qualquer outro tipo de estudo institucional.

Dessa forma, algumas propostas de melhoria e continuidade do estudo já estão sendo previstas. A primeira se refere ao aperfeiçoamento do sistema automático desenvolvido para a resolução de entidades, bem como desenvolver uma análises minuciosas de suas taxas de acertos e erros.

A segunda, e de muita importância, se refere à consolidação do novo sistema de registro da produção intelectual da USP, não mais referencial, mas sim de texto completo. Trata-se do repositório institucional intitulado Biblioteca Digital da Produção Intelectual da USP (www.producao.usp.br) que já nasceu buscando respaldo na base de recursos humanos corporativa de modo a garantir não apenas a normalização do nome do autor, mas também a identificação de várias outras características demográficas, sociais, funcionais, etc. dos autores. Tal Biblioteca Digital se vincula também com o Dedalus uma vez que deverá, doravante, fornecer e integrar os dados de produção para aquele sistema.

Como terceira estratégia, e que também se refere à estratégia mencionada anteriormente, tem-se a reativação do antigo banco de autoridade de autor visando garantir o controle de autoridade, sinonímia, nome certo, remissivas, etc.

Uma quarta proposta é o estabelecimento de norma de citação da autoria institucional USP. Tal norma deveria circular para todos os docentes USP visando orientá-los quanto à correta forma de identificação, tanto de autoria como do vínculo institucional, em suas futuras publicações. Trata-se de questão importante que terá impacto inclusive na normalização da produção USP em outras bases de indexação externas, favorecendo assim melhor e mais acurada recuperação de dados.

Como trabalho futuro, pretende-se estender a análise para as demais unidades da USP, cobrindo 100% dos autores docentes. Uma segunda vertente de ações poderá ser um estudo comparativo entre a produção USP cadastrada no Dedalus pela equipe bibliotecária e a produção cadastrada pelo próprio docente no currículo Lattes dos docentes da USP que faz parte do sistema Tycho. Esta análise permitirá verificar a consistência de diferentes tipos de dados bem como analisar sua completude.

Author names standardization in institutional information sources: a proposal for an automatic method of checking for errors

ABSTRACT

The recovery of scientific literature by authoring is a challenge for many maintainers of databases, due to the ambiguity caused by problems originated from lack of control at the time of indexing. This paper presents an automatic approach of checking for errors in authorship metadata of University of São Paulo scientific production database (Dedalus) comparing these data with data recovered from the human resources database. Using approximate string matching algorithms, these data from human resources is compared with the scientific production data of four institutes from USP (covering the period 2006-2010). Based on this pilot study it was possible to establish interoperability mechanisms between Dedalus database and the USP human resources database. As an immediate result, it was possible to map the percentage of errors and to create mechanisms of interference, establish a timeline to expand the study to other institutes from USP, and standardization procedures.

KEYWORDS: Scientific production. Authority database. Standardization. Automatic processes. Indexing.

Referências

ALCÁZAR, J. J. P. et al. Avaliação de redes de inovação usando uma ferramenta baseada em redes sociais - caso brasileiro de Nanotecnologia. In: CONGRESO LATINO-IBERO-AMERICANO DE GESTIÓN TECNOLÓGICA (ALTEC 2011), 14., 2011, Lima, Peru. **Anais...** Lima, 2011.

BILDER, G. Orcid technical update. In: ORCID TECHNICAL UPDATE COALITION FOR NETWORKED INFORMATION (CNI) ANNUAL MEETING, Fall, 2011, Arlington, VA. **Proceedings...** Arlington, VA: CNI, 2011. Disponível em: <http://www.cni.org/wp-content/uploads/2011/12/cni_orcid_bilder.pdf>. Acesso em: 25 mar. 2012.

CARVALHO, A.P. et al. Incremental unsupervised name disambiguation in cleaned digital libraries. **Journal of Information and Data Management**, Porto Alegre, v. 2, n.3, p. 289-304, 2011.

COSTAS, Rodrigo; BORDONS, María. Algoritmos para solventar la falta de normalización de nombres de autor en los estudios bibliométricos. **Investigación. bibliotecológica**, México, v. 21, n. 42, jun. 2007. Disponível em: <http://www.scielo.org.mx/scielo.php?script=sci_arttext&pid=S0187-358X2007000100002 &lng=es&nrm=iso>. Acesso em: 24 set. 2012.

CUCERZAN, S. Large-scale named entity disambiguation based on Wikipedia data. In: JOINT CONFERENCE ON EMPIRICAL METHODS IN NATURAL LANGUAGE PROCESSING AND COMPUTATIONAL NATURAL LANGUAGE LEARNING, 17., 2007, Prague. **Proceedings...** Prague: Association for Computational Linguistics, 2007. p. 708-716.

DIGIAMPIETRI, L. A.; SILVA, E. E. da. A Framework for social network of researchers analysis. **Iberoamerican Journal of Applied Computing**, Ponta Grossa, PR, v. 1, n. 1, p. 1-24, 2011

FERREIRA, A. A.; VELOSO, A. ; GONÇALVES, M. A.; LAENDER, A. H. F. Effective self-training author name disambiguation in scholarly digital libraries. In: ACM/IEEE JOINT CONFERENCE ON DIGITAL LIBRARIES, 2010, Queensland. **Proceedings…** Queensland: JCDL/ICADL, 2010. p. 39–48.

GARCÍA-GÓMEZ, Consol. Orcid: un sistema global para la identificación de investigadores. **El Profesional de la Información**, Barcelona, v. 21, n. 2, marzo/abr., 2012. Disponível em: <http://www.elprofesionaldelainformacion.com/contenidos/2012/marzo/14. pdf>. Acesso em: 21 jul. 2012.

GUERRERO-BOTE, V. et al. Method for the analysis of the uses of scientific information: the case of the University of Extremadura (1996-1997). **Libri**, Munich, v. 52, n. 2, p. 99-109, 2002.

HAN, H.; et al. Name disambiguation in author citations using a k-way spectral clustering method. In: ACM/IEEE-CS JOINT CONFERENCE ON DIGITAL LIBRARIES, 5., 2005, Denver. **Proceedings…** Denver: ACM/IEEE, 2005. p. 334–343.

KANG, I. S. et al. On co-authorship for author disambiguation. **Information Processing & Management**, Leibniz, v. 45, n. 1, p. 84–97, 2009.

MANN, G. S.; YAROWSKY, D. Unsupervised personal name disambiguation. In: CONFERENCE ON COMPUTATIONAL NATURAL LANGUAGE LEARNING (CoNLL), 7., 2003, Edmonton. **Proceedings…** Edmonton: CoNLL, 2003. p.33-40.

MÉNDEZ-VÁSQUEZ, R. I. et al. Identification and bibliometric characterization of research groups in the cardio-cerebrovascular field, Spain 1996-2004. **Revista Española de Cardiología** (English Edition), Madrid, v. 65, n. 7, p. 642–650, 2012.

ORCID: connecting research and researchers. Disponível em: <http://orcid.org/0000-0002-8205-121X>. Acesso em: 24 set. 2012.

SHIN, D. et al. Automatic method for author name disambiguation using social networks. In: IEEE INTERNATIONAL CONFERENCE ON ADVANCED INFORMATION NETWORKING AND APPLICATIONS (AINA), 24., 2010, Perth. **Proceedings…** Perth: AINA, 2010. p. 1263-1270.

TORVIK V. I. et al. A Probabilistic similarity metric for Medline records: a model for author name disambiguation. **Journal of the American Society for Information Science and Technology**, New York, v. 56, n. 2, p. 140–158, 2005.

UNIVERSIDADE DE SÃO PAULO. **Tycho**: Sistema de apoio à avaliação e a gestão institucional da USP. Disponível em: <https://uspdigital.usp.br/tycho/apresentacao.jsp?codmnu=1105>. Acesso em: 12 maio 2012.

Rogério Mugnaini
Doutor em Ciência da Informação pela Universidade de São Paulo (USP).
Professor Doutor na Escola de Artes, Ciências e Humanidades da Universidade de São Paulo (EACH/USP).
E-mail: rogerio.mugnaini@gmail.com

Luciano Antonio Digiampietri
Doutor em Ciência da Computação pela Universidade Estadual de Campinas (UNICAMP).
Professor Doutor na Escola de Artes, Ciências e Humanidades da Universidade de São Paulo (EACH/USP).
E-mail: luciano.digiampietri@gmail.com

Laucivaldo Cardoso de Oliveira
Bacharel em Ciência da Informação pela Universidade de São Paulo (USP).
Bibliotecário no Departamento Técnico do Sistema Integrado de Bibliotecas da da Universidade de São Paulo (DT/SIBi / USP)
E-mail: waldo@usp.br

Sueli Mara Soares Pinto Ferreira
Doutora em Ciências da Comunicação pela Universidade de São Paulo (USP).
Professora Titular na Faculdade de Filosofia, Ciências e Letras de Ribeirão Preto da Universidade de São Paulo (FFCLRP/USP)
E-mail: smferrei@usp.br

Recebido em: 30/09/2012

Aceito em: 07/11/2012