# FONTES DE INFORMAÇÃO INSTITUCIONAIS PARA NORMALIZAÇÃO AUTOMÁTICA DE NOMES DE AUTORES: proposta de um método automático

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

Universidade de São Paulo (USP)

Eixo temático: Bases de Dados

Modalidade: Apresentação oral

## 1 INTRODUÇÃO

A grande oferta de fontes de informação tem exigido das diversas instituições constituintes do sistema de Ciência e Tecnologia nacional uma capacidade de gestão informacional. Esta se configura uma etapa anterior a um objetivo concreto, que é a avaliação dos indicadores das diversas fontes. Porém, a concepção de um sistema de indicadores nacional que seja consistente exige a normalização de informações, desde seu registro inicial, as quais via de regra advém de inúmeros outros sistemas (ou subsistemas). Sendo os dados provenientes de sistemas distribuídos, o processo envolve distintos profissionais, e usualmente não possuem mecanismos de controle normativo que auxiliem tais profissionais a recorrer a um padrão único de entrada de nomes pessoais e institucionais. Por outro lado, além da normatização, é necessária ainda a existência de uma definição própria e única dos dados que resultarão num determinado indicador, além de sua forma correta de preenchimento.

A Universidade de São Paulo (USP) atenta à necessidade de definir indicadores de real interesse institucional, normalizados e pautados em dados confiáveis e produzidos de maneira mais dinâmica e sistêmica, criou o Grupo Permanente de Integração de Dados do Sistema Acadêmico da USP[1], 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. Dentre suas atividades pode-se citar o estabelecimento de normas de citação da autoria institucional USP nos documentos produzidos por sua comunidade, questão importante uma vez que cada membro da comunidade uspiana é o responsável pela descrição e normalização dos nomes dos autores e de suas respectivas instituições no momento de publicação de trabalhos de sua autoria.

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

Configura-se, portanto, como importante campo bibliográfico para estudo de produtividade, o campo de autoria, que apresenta o nome de autor com variações diversas, quando não incorre ainda num problema de homonímia. Esta problemática não tem sido solucionada automaticamente, ao ponto de que uma das bases de dados internacionais mais populares mundialmente, como a *Web of Science*, dá ao usuário a decisão sobre se os resultados da busca de bibliografia condizem ou não ao nome do autor buscado: oferece para isso a opção de seleção da área temática onde o mesmo costuma publicar, e as instituições de afiliação do mesmo. Se não bastasse, há que considerar as diferentes maneiras que cada um dos coautores cadastra uma mesma publicação em seus currículos na Plataforma Lattes[2] (ALCÁZAR *et al*, 2011; DIGIAMPIETRI *et al*, 2011).

Estudos diversos são encontrados na literatura com vistas à normalização de nomes de autor: desde os mais complexos (GOOI & ALLAN, 2004; GALVEZ & MOYA-ANEGÓN, 2007; COTA *et al.*, 2010), até um mais simples, que faze uso do contexto – lista de autores e departamento da universidade – para normalização dos nomes dos autores na produção científica da universidade no Science Citation Index (GUERRERO-BOTE *et al*., 2002).

Este estudo tem a pretensão de analisar comparativamente, com base especificamente na descrição dos nomes de autores, duas das bases de produção científica disponíveis na USP: o Dedalus, gerido pelo Sistema Integrado de Bibliotecas da Universidade de São Paulo (SIBi/USP), e o Sistema Thyco, que reúne os currículos Lattes da comunidade uspiana, portanto, dados inseridos pelos próprios autores.

## 2 METODOLOGIA

A metodologia está estruturada em duas partes: fontes de informação e tratamento dos dados (que inclui a obtenção e organização, e o processamento automático).

### *2.1 Fontes de informação*

#### *2.1.1 Dedalus*

Na USP, a gestão universitária dessa produção; chamada aqui de intelectual por incorporar tanto a produção científica como a acadêmica, a técnica e a artística; é desenvolvida pelo Sistema Integrado de Bibliotecas (SIBi/USP), com apoio da equipe distribuída em suas quarenta e quatro bibliotecas dispersas por diversas cidades do estado e cobrindo todas as áreas do conhecimento. A indexação dessa produção ocupa a Base 04 do

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

Dedalus: Banco de Dados Bibliográficos da USP que reúne informações sobre os diferentes acervos da Universidade, assim como proporciona dados sobre a produção intelectual. 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 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 identificado 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 3 autores e, segundo as regras do AACR2, deve ser criado uma entrada no campo 100 e outras duas no campo 700 (Entrada secundária--nome pessoal), sendo este 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 Técnicos-administrativos, mestrandos e doutorandos vinculados a USP.

### *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, pois estes demandavam levantamento de informações administrativas e acadêmicas dispersas e que se encontravam em diferentes setores da USP, como Unidades, Departamentos, Docentes, Funcionários, além de "extração ad-hoc de dados da base de dados dos sistemas corporativos" (USP, 2012), a USP desenvolve o Sistema Tycho, que apóia na avaliação e na gestão institucional da Universidade, tendo como foco a integração das atividades de coleta de dados, avaliação, diagnóstico e planejamento. (USP, 2012)

O Tycho coleta dados das seguintes fontes: base de dados corporativos existentes nos sistemas centrais mantidos pelo Departamento de Informática da Reitoria (DI) da USP, currículo Lattes e Grupos de Pesquisa do CNPq. O sistema gera grafos de colaboração, indicadores de produção bibliográfica, produção técnica, produção artística, entre outros, no período compreendido entre 1996-2012.

### *2.2 Tratamento dos dados*

#### *2.2.1 Obtenção e organização*

A base de dados do Tycho foi utilizada para a obtenção do nome dos docentes em atividade na Universidade de São Paulo. Para isso duas ferramentas diferentes foram desenvolvidas. A primeira foi utilizada para encontrar e organizar os sites correspondentes as 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. O sistema Tycho foi utilizado por conter o nome completo dos docentes, informação que nem sempre está disponível nos cadastros bibliográficos. Desta forma, 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. Vale lembrar que o Dedalus (origem dos dados do sistema Tycho) permite nomes fantasias, muito utilizados por pesquisadores das áreas de arte em geral.

A base de dados Dedalus foi utilizada para obtenção dos registros bibliográficos no formato MARC, formato que facilitou a automatização da organização e processamento dos dados. O período selecionado para estudo corresponde a 2006-2010, limitando-se a 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: identificar as diferentes formas que o nome de um docente aparece nos registros bibliográficos e verificar se existe algum tipo de inconsistência nos dados analisados.

O processamento, dividido em duas partes, buscou: identificar os autores da USP de cada registro do Dedalus na lista geral de autores (que idealmente contém um subconjunto dos nomes da lista de todos os autores das publicações, porém, os nomes destes autores podem estar armazenados de maneiras diferentes, por exemplo, sem o sobrenome de casado, com ou sem abreviações, ou ainda com ou sem alguns dos nomes do meio); identificar a correspondência entre os nomes (para cada nome na lista de autores da USP procura-se de diferentes maneiras este nome na lista total de autores). As maneiras utilizadas partem de uma

busca mais precisa até uma busca aproximada, utilizando a seguinte ordem: (i) busca exata pelo nome completo; (ii) busca exta pelo primeiro e último nome do docente e procurando por abreviações nos nomes do meio; (iii) mesma do item ii, porém permitindo a falta ou excesso do último sobrenome (caso desenvolvido para tratar o uso do "sobrenome de casado"); e (iv) busca aproximada pelo nome do docente e suas abreviações, permitindo que cada nome possua uma distância de edição de até três letras[3].

Após identificar estas correspondências entre os dados apenas do Dedalus, a mesma função de resolução de entidades foi utilizada para identificar a correspondência entre os nomes de docentes da 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 um contador de quantas vezes cada uma destas maneiras foi utilizada.

Por fim, uma função foi desenvolvida para identificar e contabilizar quais foram os principais tipos de diferenças que ocorrem nas citações dos docentes. As diferenças foram classificadas em sete categorias: último sobrenome a menos na citação, último sobrenome a mais na citação, uso de abreviações, sobrenome parecido (erro de digitação), nome parecido (erro de digitação), nomes ou sobrenomes fora de ordem, nomes a menos e outras diferenças.

## 3 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. A Tabela 1 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 (8 registros) possuem dois autores USP, mas o total de autores é apenas um, o que indica um erro no cadastramento.

Tabela 1 – Porcentagem dos registros em relação ao número de autores total e uspianos

| | | 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% |

[3] Distância de edição é a quantidade de letras que tem que ser substituídas, excluídas ou adicionadas de forma a transformar uma dada palavra em outra. Este caso foi desenvolvido para tratar de erros de digitação.

A Tabela 2 contém os diferentes tipos de produção cadastrados no Dedalus para cada uma das unidades da USP analisadas. Considerando os nomes de autores docentes registrados no Dedalus, pode-se mapear 74,2% (de um total de 1.137 docentes) 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 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 2 – 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** | |

Tabela 3 – 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 | 180 | 1.374 | *7,6* |
| Sobrenomes a menos | 35 | 282 | *8,1* |
| Abreviações | 80 | 180 | *2,3* |
| Nomes com diferenças | 23 | 165 | *7,2* |
| Nomes parecidos (erro de digitação) | 20 | 89 | *4,5* |
| Sobrenomes parecidos (erro de digitação) | 10 | 45 | *4,5* |
| Sobrenomes a mais | 7 | 37 | *5,3* |
| Nomes invertidos (ordem invertida) | - | - | - |

Tabela 4 – Ocorrências das quantidades de variação de um mesmo nome

| Quantidade de variações do nome | Total de Docentes | | Total de ocorrências | |
|---|---|---|---|---|
| | Freq. | % | Freq. | % |
| 1 | 245 | *83,6* | 7152 | *60,0* |
| 2 | 35 | *11,9* | 2608 | *21,9* |
| 3 | 11 | *3,8* | 2075 | *17,4* |
| 4 | 2 | *0,7* | 88 | *0,7* |
| 5 ou mais | - | - | - | - |

Com relação aos tipos de variação encontrados, pode-se observar na Tabela 3 que os nomes incompletos são os mais recorrentes, afetando um total de 1.374 ocorrências. Por outro lado, ao se comparar a média de ocorrências por docente, destacam-se nomes incompletos (8,1), devido à omissão de nomes/sobrenomes do meio. Alguns nomes chegam a apresentar até quatro variações, conforme observa-se na Tabela 4, por outro lado a maior parte dos nomes passíveis de normalização tem apenas uma variante.

## 4 CONSIDERAÇÕES FINAIS

Este artigo apresentou uma análise dos registros de publicações do sistema Dedalus da USP, focando nas diferentes maneiras que autores são citados na base de dados.

Para desenvolver este tipo de análise, foi necessária a identificação de uma base de referência para os nomes dos autores, neste caso, a base Thyco. Além disso, foram desenvolvidas ferramentas para a identificação e contagem automáticas das variações dos nomes, bem como para identificar potenciais problemas nos registros bibliográficos.

Como trabalho futuro, pretende-se estender a análise para as demais unidades da USP além de cruzar informações da base Dedalus com as produções cadastradas na base Thyco, as quais são extraídas automaticamente dos currículos Lattes dos docentes da USP. Esta análise permitirá verificar a consistência de diferentes tipos de dados bem como analisar a completude dos dados cadastrados na base Dedalus.

## 5 REFERÊNCIAS

ALCÁZAR, J. J. P. et al. Avaliação de Redes de Inovação usando uma ferramenta baseada em redes sociais - caso Brasileiro de Nanotecnologia. In: *XIV Congreso Latino-Iberoamericano de Gestión Tecnológica (ALTEC 2011)*, 2011, Lima, Peru. Anais do XIV Congreso Latino-Iberoamericano de Gestión Tecnológica, 2011.

COTA, R.G. *et al*. An unsupervised heuristic-based hierarchical method for name disambiguation in bibliographic citations. Journal of the American Society for Information Science and Technology, p. 61, n. 9, p. 1853–1870, 2010.

DIGIAMPIETRI, L. A., DA SILVA, E. E. A Framework for Social Network of Researchers Analysis. *Iberoamerican Journal of Applied Computing*, v. 1, n. 1, p. 1-24, 2011

GALVEZ, C., de MOYA-ANEGÓN, F. Approximate personal name-matching through finite-state graphs. Journal of the American Society for Information Science and Technology, vol. 58, n. 13, p. 1960–1976, 2007.

GOOI, C.H., ALLAN, J. Cross-document coreference on a large scale corpus. In: Dumais, S., Marcu, D., Roukos, S. (eds.) HLT-NAACL 2004: Main Proceedings, Boston, Massachusetts, USA, May 2 - May 7, pp. 9–16. Association for Computational Linguistics (2004)

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

USP. UNIVERSIDADE DE SÃO PAULO. 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: 12/05/2012.