UTILIZAÇÃO DO SOFTWARE VANTAGEPOINT NA EXTRAÇÃO DE DADOS PARA O ANUÁRIO ESTATÍSTICO DA USP

Adriana Nascimento Flamino(USP)
flamino@usp.br
Laucivaldo Cardoso de Oliveira (USP)
waldo@usp.br
Roseli Koizimi Matsuda (USP)
roselikm@usp.br
Sibele Fausto (USP)
sifausto@usp.br

EIXO TEMÁTICO: Métodos, Técnicas e Ferramentas para Estudos Bibliométricos e Cientométricos
MODALIDADE: Pôster

1 INTRODUÇÃO

O Anuário Estatístico da Universidade de São Paulo (USP) é um veículo de divulgação das mais diversas atividades das unidades da Universidade, algumas de forma genérica, outras de caráter específico, somando um total de 42 indicadores de desempenho divulgados à sociedade que a mantém e que lhe dá condições para atingir seus objetivos. A publicação do Anuário Estatístico da USP iniciou-se em 1987 com o objetivo de reunir e consolidar algumas estatísticas demográficas e acadêmicas sobre a Universidade e servir de instrumento para apoio gerencial no planejamento de suas atividades de ensino, pesquisa e de prestação de serviços (USP, 2013).

Desde 1987 o Sistema Integrado de Bibliotecas da USP (SIBiUSP) realiza a coleta de alguns dados estatísticos das bibliotecas que constam de seu Sistema, de forma padronizada e sistêmica para fornecer informações para a elaboração de tabelas que constam na publicação do Anuário Estatístico da Universidade de São Paulo. A Divisão de Gestão de Tratamento da Informação (DGTI) do Departamento Técnico do SIBiUSP (DT-SIBiUSP) gerencia o Banco de Dados Bibliográficos da USP – DEDALUS, que representa os diversos acervos das 70 bibliotecas distribuídas entre as 45 Unidades da USP. É a partir da extração de dados do DEDALUS que são disponibilizados os números da Produção Científica da Universidade, representados em tabelas e gráficos no Anuário Estatístico.

Atualmente o novo contexto de cooperação e internacionalização das Universidades (itens fudamentais para melhores posicionamentos nos diversos Rankings Universitários Mundiais), impõe que as instituições de ensino busquem indicadores que demonstrem o seu nível de participação neste novo contexto. Esses indicadores somente poderão ser apresentados a partir da extração de dados consistentes.

Soma-se a esta necessidade, no âmbito das instituições públicas, garantir a autenticidade, integridade e o acesso às informações, prevista na Lei Federal nº 12.527, de 18 de novembro de 2011 (BRASIL, 2011), a partir de uma gestão transparente da informação.

1.1 CONTEXTO DO ESTUDO

Desde 1985, a Produção Cientifica dos docentes e pesquisadores da USP vem sendo registrada e acessada por meio do DEDALUS. Após a aprovação da Resolução GR Nº 6.444, de 22 de outubro de 2012, institui-se que toda a Produção Intelectual deverá ser registrada e armazenada na Biblioteca Digital de Produção Intelectual da USP - BDPI, caracterizando uma fase de transição no gerenciamento da informação (tratamento, armazenamento, preservação) para fins de acesso à produção intelectual da USP e também para a extração de dados para fins estatísticos.

O DEDALUS utiliza o *software* proprietário ALEPH, da empresa Israelense ExLibris, e o formato de intercâmbio de dados bibliográficos MARC – *Machine Readable Cataloging*, constituído de campos e subcampos. Já a BDPI, foi desenvolvida com o *software* livre DSpace, estruturada a partir de padrões internacionais de interoperabilidade (*Open Archives Initiative – Protocol for Metadada Harvesting* - OAI-PMH) e utiliza o conjunto de metadados *Dublin Core*. Para a extração de dados estatísticos das referidas bases de dados é necessário a utilização de ferramentas específicas. Para este estudo iremos analisar a ferramenta *VantagePoint* (VP), que trabalha com dados brutos (desde que estruturados) extraídos de quaisquer bancos de dados podendo trabalhar com milhares de registros bibliográficos. Com ele é possível elaborar diversos filtros de análise e minerar dados de acordo com parâmetros pré-estabelecidos.

O VP permite processar diversos documentos com técnicas bibliométricas avançadas. Analisa a massa de dados, no caso, os registros bibliográficos que fornecemos e extrai as informações que desejamos. Segundo Porter e Palop (2012), com o VP é possível: 1) Buscar a informação nos bancos de dados textuais estruturados; 2) Baixar os resultados da busca; 3) Importar os resultados da busca para o VP; 4) Usar o VP para descobrir padrões nos resultados da busca. As funcionalidades do VP podem ser classificadas em cinco categorias: Importação: receber os dados brutos e garimpá-los para obter mais informações a partir deles; Limpeza: transformar os dados em um conjunto coerente, combinando os dados ou as informações que você deseja analisar como um grupo, fundindo e normalizando os dados de diversas fontes; Análise: explorar os dados em uma grande variedade de maneiras; Reportagem: apresentação dos resultados; Automatização: codificação de todo o processo para torná-lo consistente e facilmente reproduzível.

1.2 JUSTIFICATIVA

O Programa USP Internacional tem por objetivo fortalecer a presença da Universidade no exterior por meio de várias ações, sendo uma delas a implementação e consolidação das parcerias com instituições de ensino superior estrangeiras (A USP ultrapassa...., 2013). Nesse contexto, as informações divulgadas no Anuário Estatístico da USP por meio da antiga Tabela 9.03, publicada até 2012, com o título **"Participação de Instituições e de Coautores Externos de Outros Países, em Artigos Científicos de Periódicos Internacionais, Publicados por Docentes/Pesquisadores da USP"** são extremamente importantes e

estratégicas, pois refletem de forma muito positiva a Produção Científica (PC) da USP com a participação de coautores estrangeiros, formando uma rede de colaboração científica muito rica, diversificada e interdisciplinar.

Deste modo, a retirada dessa informação estratégica que eleva e reflete a internacionalização da USP do seu veículo de divulgação à sociedade em geral, de forma aberta e gratuita na Internet, torna-se muito preocupante, demandando soluções tecnológicas imediatas para retornarmos com esses dados ao Anuário Estatístico, seja da forma original, com os mesmos indicadores (unidade, número de Trabalhos, instituições estrangeiras, autores e países envolvidos) ou repensar os mesmos acrescentando ou retirando alguns dados.

Sendo assim, procuramos avaliar a utilização da ferramenta tecnológica VP para a análise dos dados textuais estruturados retirados do DEDALUS e, a partir desta análise, a extração dos dados estatísticos para alimentar a supracitada Tabela 9.03 do Anuário Estatístico da USP, e posteriormente demais tabelas e gráficos. Futuramente, os dados que constarão no Anuário Estatístico da USP deverão ser extraídos do DEDALUS e da BDPI. Deste modo, vários levantamentos serão necessários, mas nos atentaremos somente em uma análise preliminar, utilizando uma metodologia de uso do *software* VP quanto à extração e apresentação de dados em atendimento às demandas de informação específicas.

Este trabalho objetiva fornecer dados estatísticos consistentes sobre a rede de colaboração dos autores da USP com autores de instituições estrangeiras, como forma de contribuir para a internacionalização da USP, analisando se as informações contidas nos registros bibliográficos do DEDALUS estão devidamente estruturadas e se são passíveis de serem contabilizadas pelo VP, automatizando a extração de dados para o Anuário Estatístico da USP.

2 PROCEDIMENTOS METODOLÓGICOS

O *corpus* da análise utilizado foram os dados dos registros bibliográficos do Banco DEDALUS. Primeiramente foram realizados alguns levantamentos necessários para uma maior compreensão de como são extraídos atualmente os dados para as tabelas do Anuário.

Seguem abaixo os seguintes levantamentos:

a) Levantamento das tabelas do Anuário Estatístico da USP referentes à Produção Científica da USP, títulos e campos utilizados para a extração de dados.
b) Levantamento de tipologia documental definidos para cadastramento na Base de Produção Cientifica no Banco DEDALUS;
c) Levantamento de unidades da USP previstas para o Banco DEDALUS;
d) Levantamento de campos e subcampos utilizados nos registros bibliográficos da Base de Produção Cientifica, no Banco DEDALUS;

Após esses levantamentos, procedeu-se a uma metodologia para a extração de dados através do VP, descrita a seguir:

1. Extração de dados do DEDALUS, em seu módulo Catalog, na base de dados (04)

da Produção Científica Docente (PCD), por meio de campos e subcampos específicos, principais e secundários. Para os indicadores abaixo utilizamos os filtros elaborados para análise no VP;

2. Para o indicador 1 - total de artigos de periódicos, utilizando o campo MARC 945 subcampo $$b e também por data de publicação (de 2008 a 2012), utilizando o campo MARC 945 subcampo $$j;
3. Para os demais indicadores - filtrados por data de publicação (2012) foram adotadas as seguintes metodologias:
   Indicador 2 – Número de artigos de periódicos por unidade. Utilizamos o campo MARC 945 subcampos $$b e $$L, e o MARC 946 subcampo $$e;
   Indicador 3 – Número de artigos por Docente USP. Utilizamos o campo MARC 946 subcampos $$a e $$e, além do campo MARC 945 subcampos $$b e $$L;
   Indicador 4 – Número de autores e coautores internacionais. Utilizamos os campos MARC 100 e 700, ambos com o subcampo $$7 INT;
   Indicador 5 - Idioma dos artigos de periódicos. Utilizamos o campo MARC 041.

Os registros foram salvos em formato MARC, no bloco de notas com extensão .txt. No VP, importamos os dados. Durante esse processo, surgiram questionamentos em relação ao tratamento dos dados utilizando o VP, que são discutidos junto aos resultados obtidos.

3 RESULTADOS

As buscas no DEDALUS permitiram a recuperação de dados numéricos do seguinte recorte da produção científica da USP: n. total de artigos de periódico publicados de 2008 a 2012; n. de artigos de periódicos por unidade em 2012; n. de artigos por Docente USP em 2012; autores e coautores internacionais em 2012 e idioma dos artigos de periódicos em 2012. Os Gráficos 1, 2, 3, 4 e 5 apresentam os resultados dos levantamentos realizados, seguidos de análise e discussão.

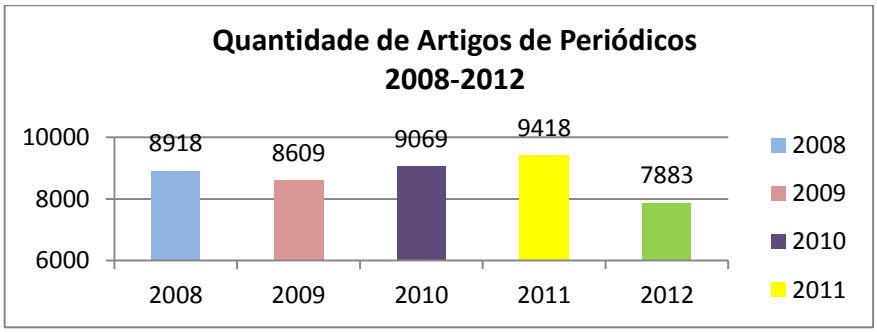

Gráfico 1: Número de artigos de periódicos 2008 a 2012

O Gráfico 1 mostra que a produção científica da USP em número de artigos de periódicos no período estudado apresentou-se com 8.918 ocorrências em 2008; 8.609 em 2009; 9.069 em 2010; 9.418 em 2011 e 7.883 em 2012. A flutuação observada pode ser condicionada a vários fatores como: a falta de cadastramento por parte dos bibliotecários; falta de entrega da produção científica do docente à Biblioteca para cadastramento;

autoarquivamento do docente ou do Publisher do artigo na Internet, seja por meio de Repositórios Institucionais, blogs, entre outros, não tendo obrigatoriedade de cadastrar no DEDALUS.

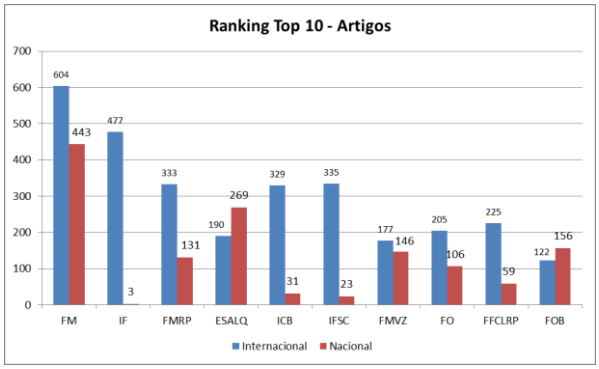

Gráfico 2: Ranking top 10 - total de artigos de periodicos por unidade em 2012

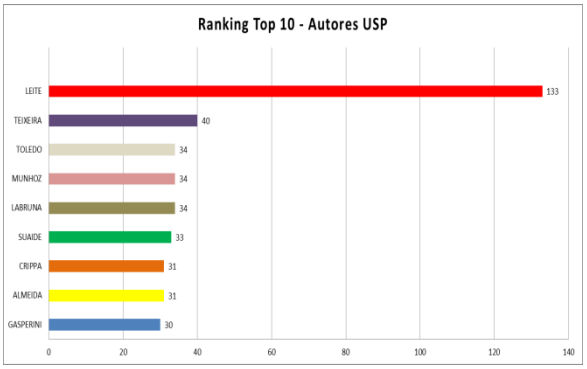

Gráfico 3: Ranking top 10 - Autores USP por número de artigos de periódicos - 2012

O Gráfico 2 mostra o total de artigos de periódicos por unidade em 2012, possibilitando um *rankeamento* das dez (10) unidades com maior número absoluto de publicações de artigos de periódico no período. Observa-se que as unidades com maior número de artigos de periódicos publicados são da área de Ciências da Saúde (Faculdade de Medicina, com 604 artigos, Faculdade de Medicina de Ribeirão Preto, com 333 artigos e Faculdade de Odontologia, com 252 artigos); Exatas (Instituto de Física, com 477 artigos, Instituto de Física de São Carlos, com 335 artigos) e a área de Ciências Biológicas (Instituto de Ciências Biomédicas, com 329 artigos e Instituto Oceanográfico, com 252 artigos); e ainda, a Faculdade de Filosofia, Ciências e Letras de Ribeirão Preto – uma unidade interdisciplinar, com 225 artigos. Verificamos que as unidades das áreas de saúde, Exatas e Biomédicas em 2012 tiveram o foco na publicação internacional, contribuindo, desta forma, para a internacionalização da pesquisa desenvolvida na USP, enquanto que o Gráfico 3 mostra o total de artigos de periódicos por docente em 2012, possibilitando observar quais docentes publicaram mais nesse período.

Confirmamos por meio de verificação de uma amostra de registros bibliográficos, que uma Unidade pode ter mais de uma ocorrência de autores no mesmo registro, mas o VP considera apenas uma ocorrência por Unidade, refletindo a consistência desse indicador.

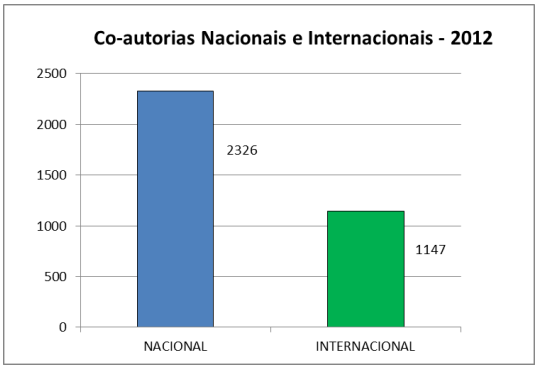

Gráfico 4: Coautores nacionais e internacionais – 2012

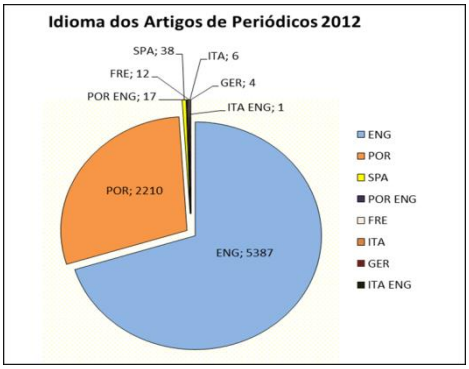

Gráfico 5: Idioma dos artigos de periódicos – 2012

O Gráfico 4 mostra o quantitativo de autores e coautores internacionais em 2012, sendo um dado imprescindível para demonstrar a rede de colaboração e acompanhar a evolução da internacionalização da USP. No entanto, verificamos que não é possível contabilizar, neste momento, a origem institucional dos autores externos, devido a não estruturação do subcampo $$8 nos campos MARC 100 e 700, tornando-se um obstáculo para o levantamento de indicadores para a alimentação da Tabela 9.03. Já o Gráfico 5 mostra os idiomas dos artigos de periódicos, sendo que o inglês é o idioma predominante, seguido do português. Outros idiomas (espanhol, francês, italiano e alemão) têm uma participação bem menor na amostra. Tal resultado reforça a constatação de que o inglês é o idioma preferencial e/ou exigido para publicação científica no mundo (MENEGHINI; PACKER, 2007). Cabe ressaltar que no contexto nacional o idioma português tem grande relevância.

4. CONSIDERAÇÕES FINAIS

O software VP é uma ferramenta para análise de dados extraídos de uma determinada fonte, e não tem por objetivo a normalização ou padronização dos dados. Portanto, se a fonte dos dados não estiver padronizada, a extração dos dados estatísticos pelo VP fica inviável e/ou inconsistente. Ressaltamos que a padronização dos dados é imprescindível para que tenhamos dados estatísticos consistentes. Sendo assim, a extração da informação "país" dos registros bibliográficos da produção científica docente do DEDALUS para alimentar a Tabela 9.03 do Anuário Estatístico da USP, no momento, não é totalmente viável, uma vez que a informação não está padronizada na fonte. Neste contexto, é preciso avaliar o custo/benefício para a padronização e estruturação do subcampo $$8 dos campos MARC 100 e 700 para a efetiva extração dos dados, uma vez que estamos em processo de transição de fontes de informação, ou seja, do DEDALUS para a BDPI. Para essa atividade, o DT-SIBiUSP precisaria envolver todas as bibliotecas do Sistema para efetivar a padronização dos registros retrospectivos, uma vez que essa padronização via máquina não é viável. Por exemplo, muitos registros não tem a informação de país, ou estão fora de ordem, ou possuem a informação da cidade ao invés de país, entre outros, demandando verificação individual de cada registro.

Sendo assim, o DT-SIBiUSP precisa definir a fonte (ou as fontes) de informação corporativa oficial e planejar metodologias e estratégias de padronização e extração dos dados. Definir se a alteração nos registros será somente do ano de 2013 visando à extração correta dos dados para o Anuário de 2014, ficando a critério das bibliotecas a alteração dos anos retrospectivos mediante as condições, necessidades e demandas de cada uma delas ou se a alteração será de todos os anos e obrigatória para todas.

Portanto, a definição de novas políticas e diretrizes por parte do DT-SIBiUSP é fundamental para a padronização dos dados no DEDALUS e BDPI para a geração de dados consistentes para a apresentação em relatórios gerenciais internos e externos, relatórios demandados pela Reitoria da USP, INEP, Anuário Estatístico da USP, entre outros. Aumentando, desta forma, a credibilidade do DT-SIBiUSP na qualidade e consistência das informações divulgadas, no gerenciamento eficaz e efetivo do acervo das bibliotecas do

Sistema, da produção científica da Universidade, no uso das fontes de informações adquiridas pelo DT-SIBiUSP, na prestação de serviços a comunidade USP e à sociedade em geral, contribuindo assim, de forma concreta para a internacionalização da USP.

Destaca-se que não foi objetivo deste trabalho levantar críticas ou problemas relacionados ao desempenho do *software VantagePoint*, mas sim, foi um exercício inicial, fruto do Trabalho de Conclusão de Curso de Capacitação de Bibliotecários em Análise Bibliométrica para Apoio à Gestão da Pesquisa em Universidade Pública, no âmbito da parceria SIBiUSP e Escola Técnica e de Gestão da USP, visando obter parâmetros e subsídios para futuros desenvolvimentos e customizações do aplicativo para atender a diversas demandas do SIBiUSP com a aquisição do referido software.

REFERÊNCIAS

A USP ultrapassava fronteiras**. Boletim USP Destaques**, n. 71, 09 Abr. 2013. Disponível em: http://www.usp.br/imprensa/wp-content/uploads/USP-Destaques_71.pdf. Acesso em 28 abr. 2013.

BRASIL. Lei n.º12.527, de 18 de novembro de 2011. Regula o acesso a informações previsto no inciso xxxiii do art. 5º, no inciso ii do § 3º do art. 37 e no § 2º do art. 216 da constituição federal; altera a lei nº 8.112, de 11 de dezembro de 1990; revoga a lei nº 11.111, de 5 de maio de 2005, e dispositivos da lei nº 8.159, de 8 de janeiro de 1991; e dá outras providências. **Diário Oficial da União,** 18/11/2011, p. 1(Edição Extra). Disponível em: http://www.planalto.gov.br/ccivil_03/_ato2011-2014/2011/lei/l12527.htm. Acesso em 17 mai. 2013;

MENEGHINI, R.; PACKER, A. L. Is there science beyond English? Initiatives to increase the quality and visibility of non-English publications might help to break down language barriers in scientific communication. **EMBO Rep**. v. 8, n. 2, p. 112–116, feb. 2007.

PORTER, A.; PALOP, F. Mineração de textos para decisões de gestão de pesquisa e tecnologia - tech mining com a ajuda de software Vantage Point. In: Encontro Brasileiro de Bibliometria e Cientometria, 3., Gramado, RS. Anais... Gramado: UFRGS, 2012. Disponível em: http://www.ufrgs.br/ebbc2012/arquivos/workshop-2. Acesso em: 01 Mai.2013

UNIVERSIDADE DE SÃO PAULO (USP)**. Anuário Estatístico**: Apresentação. Disponível em: https://uspdigital.usp.br/anuario/apresentacao.jsp?codmnu=2781 . Acesso em: 28 Abr.2013.

_____. Reitoria. Resolução nº 6444, de 22 de outubro de 2012. Dispõe sobre diretrizes e procedimentos para promover e assegurar a coleta, tratamento e preservação da produção intelectual gerada nas Unidades USP e pelos Programas Conjuntos de Pós-Graduação, bem como sua disseminação e acessibilidade para a comunidade**. Diário Oficial do Estado**, 23 de out. 2012a. Disponível em: http://www.usp.br/drh/novo/legislacao/doe2012/res-usp6444.html. Acesso em 10 Abr. 2013.

_____. Reitoria. Portaria GR-5917, de 22 de outubro de 2012. Altera dispositivo da Portaria GR-2.922/1994, de 16 de novembro de 1994. **Diário Oficial do Estado**, 23 out. 2012b. http://www.leginf.usp.br/?portaria=portaria-gr-no-5917-de-22-de-outubro-de-2012. Acesso em 04 mai. 2013.

_____. Reitoria. Portaria GR-2922, de 16 de novembro de 1994. Regulamenta o

funcionamento do Banco de Dados Bibliográficos da Universidade de São Paulo e dá outras providências correlatas**. Diário Oficial do Estado**, 18 nov.1994. Seção I, p. 58. Disponível em: http://citrus.uspnet.usp.br/sibi/Portaria-Resolucao/port_gr_2922.htm . Acesso em: 22 dez. 2012.