Desenvolvimento de um sistema de informação para coleta e análise de relatórios de produtividade científica

Beneficiário: Bruno Boaventura Scholl[1,2]
Orientador: Marcelo S. Reis[1]

[1] Center of Toxins, Immune-response and Cell Signaling (CeTICS);
Laboratório Especial de Ciclo Celular (LECC), Instituto Butantan
[2] Instituto de Matemática e Estatística (IME), Universidade de São Paulo (USP)

São Paulo, 19 de abril de 2019

Resumo

Avaliações de produtividade são parte essencial do monitoramento e controle da produção do conhecimento científico. A Divisão de Desenvolvimento Científico (DDC) do Instituto Butantan elabora anualmente um detalhado relatório de produtividade, baseado em informações fornecidas pelos pesquisadores dos diversos laboratórios do Butantan. Todavia, os processos atuais de produção desse relatório, a despeito de avanços recentes, demandam muito tempo e trabalho tanto por parte dos pesquisadores quanto do comitê nomeado pela DDC para compilar e analisar esses dados. Dessa forma, neste projeto propomos resolver essa questão através do desenho e implementação de um banco de dados relacional, populando-o com dados do relatório atual e também os de edições anteriores. Numa segunda etapa, analisaremos esses dados através de consultas ao banco de dados e também com a elaboração de grafos de colaboração entre pesquisadores. Por fim, desenvolveremos uma interface web para que as próximas edições do relatório da DDC sejam coletadas de forma rápida e automatizada. Ao final do projeto, o código-fonte será disponibilizado de forma livre e gratuita, o que permitirá que o mesmo também possa ser empregado em outros Institutos de Pesquisa do Estado de São Paulo.

Palavras chave: Cientometria, Bibliometria, Sistema de informação, Banco de dados

Sumário

1 Introdução

Avaliações de produtividade são mecanismos através dos quais a comunidade acadêmica, de forma periódica, monitora e controla a produção do conhecimento científico. Essas avaliações são utilizadas pelas lideranças institucionais para diversos fins, tais como a concessão de bolsas e auxílios e a contratação e/ou promoção de pessoal [1]. No caso particular do Instituto Butantan, uma das mais importantes avaliações de produtividade é a organizada anualmente pela Divisão de Desenvolvimento Científico (DDC). Os resultados obtidos nessa avaliação são enviados à Secretaria da Saúde do Estado de São Paulo, como forma de prestação de contas à sociedade paulista, e também utilizados internamente para definição de políticas científicas e distribuição de recursos entre os laboratórios do Butantan.

Tradicionalmente o relatório anual da DDC é elaborado em duas etapas: na primeira delas, pesquisadores de cada laboratório preenchem dois formulários. O primeiro deles é um documento Word no qual são inseridas informações tais como lista de publicações de artigos e de livros ao longo do ano, orientações concluídas e em andamento, participação em congressos, aulas ministradas, etc. Já o segundo, implementado em uma planilha Excel, serve especificamente para colocar dados financeiros de auxílios e de bolsas de alunos. A primeira etapa é concluída quando a secretaria de cada laboratório compila os dados de todos os pesquisadores em um único documento Word e uma única planilha Excel; ou seja, são produzidos dois arquivos que são referentes ao laboratório como um todo. Já na segunda etapa, também conhecida como *consolidação do relatório*, um comitê nomeado pela DDC recebe todos os relatórios de laboratório (atualmente 29 deles, de quatro divisões diferentes, incluindo a própria DDC) e gera estatísticas holísticas sobre os mesmos: por exemplo, número de pesquisadores do Butantan (distinguindo entre estatutários e celetistas contratados pela Fundação Butantan), número de alunos orientados pelos pesquisadores (separando entre bolsistas e não bolsistas), número trabalhos publicados (distinguindo entre artigos, livros e capítulos de livros), total de valores arrecadados em auxílios e bolsas (separando por valores em real e dólar), etc. Como todo esse processo é feito de forma manual, a consolidação do relatório é um processo muito trabalhoso, demorado e sujeito a erros na compilação das informações.

Recentemente, o comitê atualmente designado pela DDC realizou algumas melhorias no processo de elaboração da edição de 2018 do relatório. Uma delas foi a simplificação da documentação, que passou a ser composta por apenas uma planilha Excel para todas as informações, facilitando assim o preenchimento da mesma por parte do pesquisador; tal simplificação também eliminou redundâncias que existiam entre os dois arquivos da documentação original. Outro avanço importante foi na consolidação do relatório: como o novo formulário foi criado com a preocupação em manter um posicionamento sistemático das informações (e.g., nome de aluno bolsista só pode ser inserido na coluna B, entre as linhas 30 e 70), para realizar a consolidação bastou ao comitê inserir num mesmo arquivo todos formulários de laboratórios, um por aba; após isso, foram incluídas abas que "puxam"as informações das abas dos laboratórios para gerar as estatísticas necessárias; com o posicionamento sistemático das informações, as fórmulas nas abas de estatísticas "sabem"onde estão localizadas as informações necessárias.

Todavia, embora o relatório de 2018 tenha tido os avanços descritos acima, ainda não é possível fazer de forma imediata e abrangente uma comparação desses resultados com os obtidos nos anos anteriores: seria preciso fazer antes uma curadoria manual da documentação das edições passadas, que está sujeita a todos os problemas mencionados anteriormente, e que portanto precisa ser feita com tempo e cuidado. Além disso, planilhas tais como as do Excel não são o meio mais adequado para armazenar e analisar o grande volume de informações gerado pela série histórica de relatórios: uma alternativa escalável, e portanto mais apropriada para esse fim, seria um banco de dados relacional. Por fim, embora o formulário único tenha sido um avanço importante em relação ao relatório de anos anteriores, uma interface web cobrindo as mesmas informações seria uma alternativa que facilitaria ainda mais o preenchimento por parte do pesquisador; ademais, se ao mesmo tempo essa interface alimentar um banco de dados relacional, o processo de consolidação do relatório anual da DDC tornar-se-ia totalmente automatizado, aliviando assim o comitê desse relatório.

2 Objetivos

Este projeto proposto tem três objetivos principais:

1. Desenhar um banco de dados relacional para acomodar os dados dos relatórios de produtividade DDC das edições de 2018, 2017 e, se possível, também de outras edições anteriores;
2. Fazer a avaliação de produtividade propriamente dita, utilizando para isso consultas ao banco de dados do item anterior. Serão feitas análises tanto do ponto de vista dos laboratórios quanto dos pesquisadores individualmente;
3. Desenhar e implementar uma interface web amigável para auxiliar na elaboração de relatórios de produtividade futuros, armazenando as informações diretamente no banco de dados mencionado acima.

Espera-se que, ao final deste projeto, tenhamos um sistema de informação para coleta e análise de relatórios de produtividade científica que, embora tenha sido originalmente desenhado para a elaboração do relatório anual de produtividade DDC, também possa ser utilizado em outros Institutos de Pesquisa do Estado de São Paulo.

3 Metodologia

Os três objetivos principais deste projeto proposto podem ser organizados como três sub-projetos, um sequencial ao outro, cada qual com suas próprias técnicas e ferramentas. A seguir, detalharemos a metodologia que será empregada em cada um desses sub-projetos.

3.1 Desenho e população do banco de dados relacional

O banco de dados relacional será inicialmente projetado tomando como base as tabelas existentes no relatório anual da DDC de 2018, eventualmente complementada por critérios cientométricos

utilizados em avaliações recentes das universidades estaduais paulistas [2]. O banco de dados será desenhado utilizando modelo entidade-relacionamento e implementado com o sistema gerenciador de banco de dados relacional MySQL [3].

Para a população do banco de dados, as abas das planilhas Excel do relatório de 2018 poderão ser exportadas em formato `csv` e processadas com programas que filtram o arquivo e montam os comandos de inserção de informação no banco de dados. Uma linguagem de programação adequada para este fim é a Perl [4], que conta com amplo suporte a expressões regulares, algo muito útil para o processo de filtragem dos arquivos `csv`. No caso dos relatórios de anos anteriores, o fato das informações contidas nos mesmos não contarem com um nível suficiente de sistematização exigirá que os comandos de inserção no banco de dados sejam elaborados de forma manual.

3.2 Avaliação dos relatórios de produtividade

A avaliação inicial do relatório será feita através de consultas SQL ao banco de dados, que serão feitas de acordo com as estatísticas que são necessárias para a elaboração da versão final do relatório anual de produtividade DDC. Um programa em Perl pode fazer automaticamente as consultas e gerar um arquivo `csv`, que por sua vez pode ser importado para dentro de uma planilha Excel. Por fim, os dados dessa planilha podem ser utilizados para gerar gráficos, utilizando para este fim as próprias ferramentas que o Excel possui.

Numa segunda etapa, será feita pela primeira vez pela DDC uma avaliação individual anonimizada dos pesquisadores (i.e., na divulgação da avaliação os nomes dos pesquisadores serão substituídos por códigos aleatórios), nos moldes do que já foi feito pelo Departamento de Química Fundamental do Instituto de Química da USP [5]. Essa avaliação será feita utilizando critérios cientométricos que foram incorporados ao relatório anual em sua última edição. Mapearemos também o nível de colaboração existente entre os pesquisadores do Instituto Butantan, montando e analisando grafos de colaborações. Para desenhar tais grafos, inicialmente utilizaremos a ferramenta ScriptLattes [6]. Nossa ênfase será na análise dos laboratórios que atualmente são fisicamente isolados e que em breve compartilharão o mesmo espaço [1] – o antigo prédio de Recursos Humanos do Butantan deverá ser reformado e empregado para esse fim. Uma das expectativas existentes com essa mudança é de que surjam sinergias não somente no uso de equipamentos, mas principalmente nas colaborações entre pesquisadores de diferentes laboratórios: a evolução do grafo de colaboração nos próximos anos será uma boa referência para testar essa hipótese.

As atividades deste subprojeto serão acompanhadas de perto pelo orientador deste projeto proposto, membro do comitê de elaboração do relatório deste ano, e também pelos demais membros do comitê: Dra. Sandra Vessoni (Diretora da DDC), Dra. Maria Carolina Elias (Diretora-substituta da DDC), Dra. Julia Cunha (pesquisadora do LECC), Dr. Enéas Carvalho (pesquisador do Laboratório de Bacteriologia), Joanita Lopes (Diretora Técnica da Biblioteca

[1] Os laboratórios envolvidos e o cronograma de mudança ainda serão definidos pela direção do Instituto Butantan, pela DDC e por lideranças científicas da instituição.

do Butantan) e Sarah Oliveira (Supervisora da Biblioteca do Butantan). A avaliação será feita procurando seguir boas práticas de bibliometria científica [7].

3.3 Desenho e implementação da interface web

Neste último sub-projeto desenvolveremos uma interface web para substituir a atual planilha Excel que é preenchida por cada um dos pesquisadores do Butantan. O sistema deverá ser projetado utilizando a técnica de casos de uso e notação UML [8], novamente tomando como ponto de partida a planilha Excel atual.

Embora ainda estudaremos qual seria a melhor tecnologia a ser empregada na implementação da interface, provavelmente deveremos utilizar o arcabouço Django [9]. Procuraremos seguir boas práticas de programação, tais como controle de versão do código, testes unitários e de integração, e disponibilização do código-fonte final de forma livre e gratuita, utilizando para isso uma licença tal como a GNU-GPL.

4 Plano de trabalho

Este projeto proposto será executado ao longo de 2019 pelo beneficiário, no contexto da disciplina do IME-USP "Trabalho de Formatura" (MAC499). Essa disciplina exige que, no final do ano, o aluno entregue uma monografia, elabore e apresente um pôster e, opcionalmente, também faça uma apresentação oral. Obviamente, informações confidenciais presentes no banco de dados não serão apresentadas nesses materiais.

Pela estrutura sequencial desta proposta em três sub-projetos, o cronograma de execução é relativamente simples; o mesmo é apresentado abaixo em dois diagramas de Gantt mostrados nas tabelas 1 e 2.

4.1 Cronograma de execução

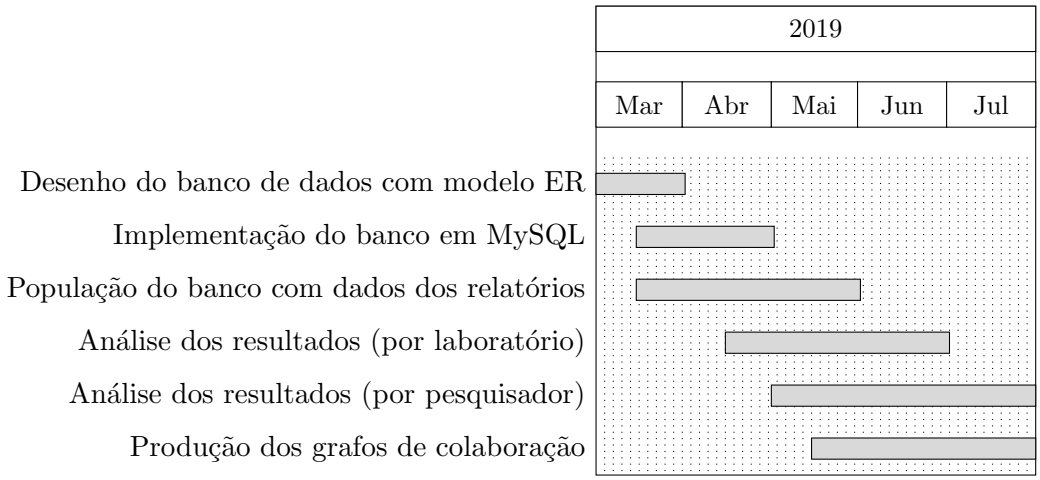

Tabela 1: Cronograma de execução dos sub-projetos 1 e 2 deste projeto proposto.

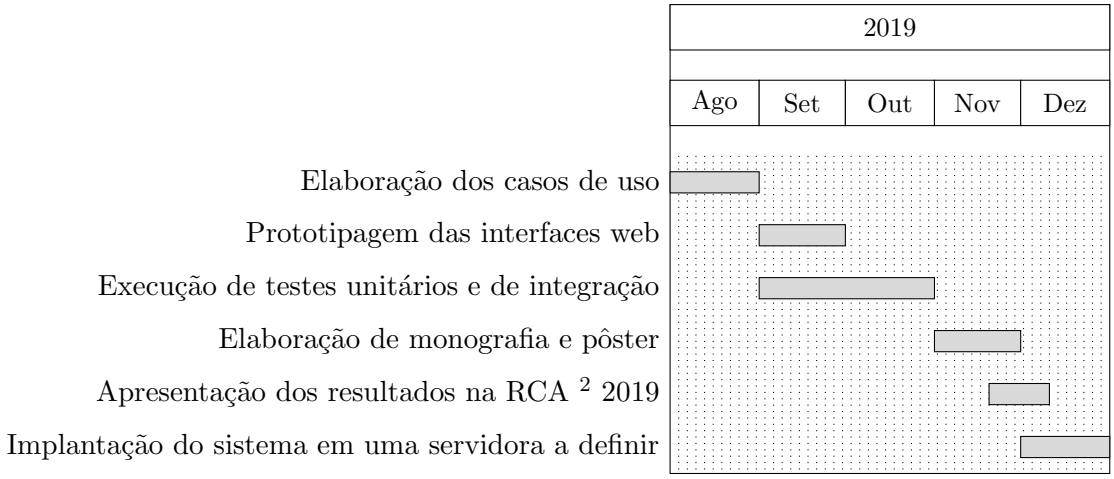

Tabela 2: Cronograma de execução do sub-projeto 3 deste projeto proposto.

[2]RCA: Reunião Científica Anual do Instituto Butantan.

Referências

[1] Marcos Antônio Mattedi and Maiko Rafael Spiess. A avaliação da produtividade científica. *História, Ciências, Saúde-Manguinhos*, 24(3), 2017.

[2] Jacques Marcovitch et al. *Repensar a universidade: desempenho acadêmico e comparações internacionais*. ComArte, 2018. Livro digital disponível gratuitamente em: http://www.livrosabertos.sibi.usp.br/portaldelivrosUSP/catalog/book/224.

[3] Paul DuBois and Michael Foreword By-Widenius. *MySQL*. New riders publishing, 1999.

[4] Randal L Schwartz and Tom Phoenix. *Learning Perl*. O'Reilly & Associates, Inc., 2001.

[5] Fabrício Marques. O melhor de cada um, 2015. revistapesquisa.fapesp.br/2015/06/16/o-melhor-de-cada-um/.

[6] Jesús Pascual Mena-Chalco and Roberto Marcondes Cesar Junior. ScriptLattes: an open-source knowledge extraction system from the Lattes platform. *Journal of the Brazilian Computer Society*, 15(4):31–39, 2009.

[7] Diana Hicks, Paul Wouters, Ludo Waltman, Sarah De Rijcke, and Ismael Rafols. Bibliometrics: the Leiden manifesto for research metrics. *Nature News*, 520(7548):429, 2015.

[8] Perdita Stevens and Rob Pooley. *Using UML: software engineering with objects and components*. Addison-Wesley Longman Publishing Co., Inc., 1999.

[9] Adrian Holovaty and Jacob Kaplan-Moss. *The definitive guide to Django: Web development done right*. Apress, 2009.